IntechOpen

Public Health

Methodology, Environmental and Systems Issues

Edited by Jay Maddock

PUBLIC HEALTH – METHODOLOGY, ENVIRONMENTAL AND SYSTEMS ISSUES

Edited by **Jay Maddock**

Public Health - Methodology, Environmental and Systems Issues
http://dx.doi.org/10.5772/2678
Edited by Jay Maddock

Contributors

Majid Rezaei Basiri, Mahmoud Ghazi-Khansari, Hasan Rezazadeh, Iraj Aswadi-Kermani, Mohammad Ali Eghbal, Alireza Partoazar, Irina Denisova, Marina Kartseva, Dan I. Riga, Pascale Allotey, Daniel Reidpath, Shajahan Yasin, Octavio Gómez-Dantés, Julio Frenk, Conrad Iyegbe, Margarita Rivera-Sanchez, José María Gutiérrez, Yuichiro Abe, Nathan Grills, Fernando Luiz Pereira De Oliveira, Luiz Duczmal, Anderson Duarte, Andre Cancado, Marcus Vinicius Teixeira Navarro, Handerson Jorge Dourado Leite, Christopher Charles, Lisa Campo-Engelstein, Sarah Rodriguez, Shauna Gardino, María Constanza Lozano, Mary Trujillo, Ulrike Ravens-Sieberer, Veronika Ottova, Anders Hjern, Carsten-Hendrik Rasche, Pawel Kretowicz, Tomasz Chaberko, Francisco De Assis S. Santos, Renato Garcia, Salvatore Guglielmino, Marco Sebastiano Nicolò, Sean Curran, Charles J. Brumlik, Steve Sund, Kwok Wai Lem, Abhishek Choudhury, Dai Soo Lee, Shan-Hui Hsu, Zafar Iqbal, David S-G Hu, Nelson Chiu, Richard Lem, Jung Rim Haw, Houngbo P. Thierry

Notice

First published in Croatia, 2012 by INTECH d.o.o.
eBook (PDF) Published by IN TECH d.o.o.
Place and year of publication of eBook (PDF): Rijeka, 2019.
IntechOpen is the global imprint of IN TECH d.o.o.
Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from orders@intechopen.com

Public Health - Methodology, Environmental and Systems Issues
Edited by Jay Maddock

p. cm.

ISBN 978-953-51-0641-8

eBook (PDF) ISBN 978-953-51-7007-5

Meet the editor

Professor Jay Maddock, Ph.D., FAAHB is Professor and Director of the Office of Public Health Studies at the University of Hawai'i at Mānoa and Luojia Chair Professor at Wuhan University. Professor Maddock has extensive experience in system, environmental and policy research to improve population level risk factors for chronic disease. Dr. Maddock chaired the state board of health and was a charter member of the NIH study section on Community-Level Health Promotion. He is an author of over 75 scientific articles. He is the Honorary Secretary for the Asia-Pacific Academic Consortium for Public Health. His research has been featured in several national magazines including Eating Well, Prevention and Good Housekeeping.

Contents

Preface

Public health can be thought of as a series of complex systems. Many things that individual living in high income countries take for granted like the control of infectious disease, clean, potable water, low infant mortality rates require a high functioning systems comprised of numerous actors, locations and interactions to work. Many people only notice public health when that system fails. With widespread globalization occurring, public health issues have become transnational. Infectious diseases like SARS, H1N1 or the common cold can be transmitted within hours across national borders via airplane. Pollution and environmental degradation can be outsourced from high income countries to lower income countries via trade imbalances in manufacturing or recycling. Even NCDs can be transmitted via the global market for tobacco and fast food. For public health to continue to protect the public from these threats clear systems thinking with the development of novel methodologies is needed.

The first section of this book explores novel measurement and methodologies for a variety of public health concerns. Chapters include assessing risk and uncertainty, measurement of mental health in children, the use of the Ames assay and measuring gene by environment interactions. The second section examines issues in the food system and environmental risks. A safe, reliable food system is essential for public health in every country. Issues in this section include the presence of chemical residues in animal food products, bacteria in food and iron deficiency anemia. The two environmental health chapters include snakebites, one of the oldest public health problems and waste minimization in nanosilver productions one of the newest public health concerns. The third section of the book reviews some of the major challenges in health systems. These include health resources, technology and management of medical devices. The role of private business in public health is also explored. The final section contains a variety of issues related to global health. This includes the rise of NCDs in low and middle income countries, neglected diseases related to poverty and health and longevity medicine. A chapter of alcoholism and mortality examines the effects of a public health system breakdown. Final chapters review men's health, insomnia and a gendered analysis.

This book exemplifies the global nature of public health. All six inhabited continents are represented by authors in this book. The home country of the authors include

Australia, Turkey, Poland, Mexico, Brazil, Canada, Korea, The Netherlands, Japan, Benin, Malaysia, USA, Russia, Romania, Taiwan, Iran, Costa Rica, Columbia, Sweden, Germany and Italy. This trans-national list of authors provides an important view of the future of public health and the increased need to collaborate with public health professionals across the world to address the myriad of public health issues. I hope you enjoy reading the following chapters. I find them to be insightful and to provide an excellent collection of the ways that methodology advances and systems sciences are being used to protect and promote the public's health. Aloha.

Prof. Jay Maddock
Department of Public Health Sciences,
University of Hawai'i at Mānoa
USA

Section 1

Measurement and Methodology

1

Potential Risk: A New Approach

Handerson J. Dourado Leite and Marcus V. Teixeira Navarro
Federal Institute of Education, Science and Technology of Bahia
Brazil

1. Introduction

Risk is a polysemic term that has been transformed throughout the historical process, but has always been associated to the idea of predicting an unwanted future event.

The first rudimentary notion of what can be called risk, may have arisen, according to Covello and Munpower (1985), around 3200 BC in the valley between the Tigris and Euphrates Rivers, where lived a group called "*Asipu*". A major function of this group was to help people who needed to make difficult decisions. The "*Asipus*", when sought, identified the scale of the problem, the alternatives and the consequences of each alternative. Then, they drew up a table, marking the positive and negative points of each alternative to indicate the best decision.

With the great voyages in the fifteenth century it became necessary to evaluate the damage caused by the potential loss of ships. Emerges then the term risk, with connotations similar to what is meant today, but the understanding of its causes was related to accidents and, therefore, impossible to predict. The development of classical probability theory, in the mid-seventeenth century, to solve problems related to gambling, allowed the start of the process of quantifying the risks, but the causes were still credited to chance.

Only from the nineteenth century, associated with the dominant thinking of the primacy of science and technique and propelled, among other factors, by the discoveries of Pasteur, emerged the association of risk with prevention, i.e., if the causes are known and quantified one can predict the undesirable effects.

The advent of modernity has produced and incorporated to the human way of life a variety of technologies and the risk became the distinguishing feature of this generated complexity. More and more, the sources of hazards[1] were associated with daily social practices. In today's society, it is difficult to separate the manmade dangers of the "natural" dangers (Beck, 2003). A flood for example, that occurred as a completely spontaneous phenomenon, today can happen as a consequence of human action on nature. This new concept that the term risk assumes defies the human prediction capacity and rationality, because its causes are no longer accidental and the causes are not always known, or they are possible effects of the technologies generated by man himself.

[1] Hazards are "physical, chemical or biological agents or a set of conditions that present a source of risk." (Kolluru, 1996. p. 3-41).

2. Risk and probability

The first report of a quantitative risk evaluation applied to health goes back to Laplace, in the late eighteenth century, which calculated the probability of death among people with and without vaccination for smallpox. With Pasteur's studies in the late nineteenth century, it was possible to use the tools of statistics to evaluate the factors related to communicable diseases, giving birth to the concept of epidemiological risk (Covello; Munpower, 1985, Czeresnia, 2004).

Epidemiological studies about contagious diseases have two very specific characteristics. The first refers to the object, which is only a source of damage. The second relates to the goals, which aim to determine the relationship between cause and effect, i.e., between exposure and disease. So, even with multifactorial determinants, it's an unidimensional evaluation. Therefore, in a evaluation between exposed and unexposed, the concept of risk approaches the definition of probability. However, when the objective includes the judgment about the severity of the injury or the comparison of different injuries in different exposures, the probability becomes one of the information that compose the concept of risk.

Therefore, the development of probability enabled the start of the process of quantifying risk. However, it's noteworthy that probability and risk are different concepts to most subjects. While the probability it's mathematically defined as the possibility or chance of a particular event occurs, and is represented by a number between 0 and 1 (Gelman; Nolan, 2004, Triola, 2005), the risk is associated with the probability of occurrence of an undesired event and its severity and cannot be represented by only one number.

If two events A and B have, respectively, 0.10 and 0.90 probability of occurring, the event B is classified as nine times more likely to occur than the event A. However, one can not say that the event B has a greater risk that the event A. For the concept of risk, is fundamental to know how much the event will be harmful. The evaluation of the probabilities of occurrence of the events A and B is done purely with mathematical analysis, while the risk assessment requires judgment of values. Thus, all observers will agree that the event B is more likely to happen than the event A, but not all should agree on which event represents a greater risk, knowing, or not, the damage.

As already explained, the notion of risk has been transformed throughout human history, it being understood nowadays as a theoretical elaboration that is historically constructed in order to mediate the relationship between man and the hazards, in order to minimize losses and maximize the benefits. Thus, it is not a greatness that is in nature to be measured, is not independent of the observer and his interests. It is formulated and evaluated within a political-economical-social context, having a multidimensional and multifactorial character (Fischhoff et al., 1983, Covello; Munpower, 1985, Beck, 2003, Hampel, 2006)

3. The risk in the modern era

The beginning of the twentieth century was marked by great scientific advances. The application of this knowledge produced new technologies such as X-rays, nuclear energy, asbestos and formaldehydes. The rapid use of these technologies as if they were only sources of benefits brought consequences to public health and to the environment, which only came to be perceived and understood by society, from the 70s of the last century. The disclosure of these risks led to pressures on governments, to control occupational,

environmental, chemical agents and radioactive agents risks. In this context of large social movements, the need for State intervention was strengthened, in order to regulate the use of products potentially harmful to health and the environment (National Research Council, 1983, Lippmann; Cohen; Schlesinger, 2003, Omenn; Faustman, 2005)

The regulation of health risks is understood as a government interference in the market or in social processes, in order to control potentially damaging consequences to health (Hood; Rothstein; Baldwin, 2004). The model of the regulatory system, deployed in each country depends on political, economic and social conjunctures. Therefore, in the 1970s, while European countries exerted, initially, its regulatory power, by means of direct administration bodies of the State, the United States exercised this power, mainly, through independent and specialized agencies.

Currently, most European Union countries use the model of regulatory agencies (Lucchese, 2001). In Brazil, this role it's exercised in a hybrid way, because the National System of Sanitary Surveillance (*Sistema Nacional de Vigilância Sanitária* - SNVS) is composed of a regulatory agency in the federal sphere, the National Health Surveillance Agency (*Agência Nacional de Vigilância Sanitária* – ANVISA), but in most states and municipalities the regulation is exerted by direct administration.

The new technologies permeate the entire society and, therefore, influence and change the established social relations. These technologies are characterized by having intrinsic risks, by the possibility of adding new risks throughout their life cycle and by the incomplete scientific knowledge about the types of risks they generate and their interactions in different situations. Thus, the regulatory process occurs, in most cases, in situations of epistemic uncertainty, where risk factors are presented in a diffuse way, requiring from sanitary surveillance the use of mutually complementary strategies of health protection.

As for the economic and social consequences related to the decisions of regulatory actions were amplified by the globalization process, as many decisions go beyond national borders and bring into play great interests. The first regulatory decisions showed that the process of definition and regulation of risk is an exercise of power, full of interests and political, economical, and social concepts, and can strongly influence the allocation of public and private resources of a nation (Slovic, 2000, Fischhoff; Bostrum e Quadrel, 2005).

Thus, the risk conceived as the probability of occurrence of an undesired event, calculated by specialists and presented to society as an absolute and neutral truth, began to be questioned. The conflicts of interest over the division of risk showed that it is not possible to separate the technical analysis about the risks from the decisions of who should be protected, from the costs and from the available alternatives, because the studies or risk evaluations occur, necessarily, to subsidize decision-making.

4. Other dimensions of risk

The fact that the calculation of risks undertaken by experts no longer represented the absolute truth and, also, the impossibility to eliminate the risks produced by the new technologies, because the benefits would also be suppressed, bring up new angles for the analysis of the phenomenon. Therefore, come into play other dimensions of risk as acceptability, perception and confidence in the regulatory system.

In beginning of the 1980, the U.S. Congress, realizing the need to structure a model of risk assessment that had wide acceptance, as well as standardizing the realization of studies in various areas, established a directive that designated the Food and Drug Administration (FDA) as responsible in coordinating a study for the harmonization. The FDA commissioned the National Academy of Sciences of the United States, which developed the project, whose results were of notorious and acknowledged importance, structuring the foundation for the paradigm of risk regulation (National Research Council, 1983, Omenn, Faustman, 2005).

This study, published in 1983 under the title *Risk assessment in the government: managing the process*, known internationally as the *Red Book*, establishes a process with seven stages: (1) Hazard identification, (2) dose x response assessment, (3) exposure assessment, (4) risk characterization; (5) Establishment of regulatory options, (6) Decision and implementation of the option of regulation, (7) Evaluation of the regulation. All steps occur with the participation of various actors, experts or not. The stages (1 to 4) are classified as risk assessment and are of technical and scientifically base. The other stages (5 to 7) are part of risk management, which, taking into account the information obtained in the first stage, evaluate and implement the best regulatory options, considering economical, political and social issues.

A diagram of the paradigm of risks applied to the area of health surveillance is represented in Figure 1.

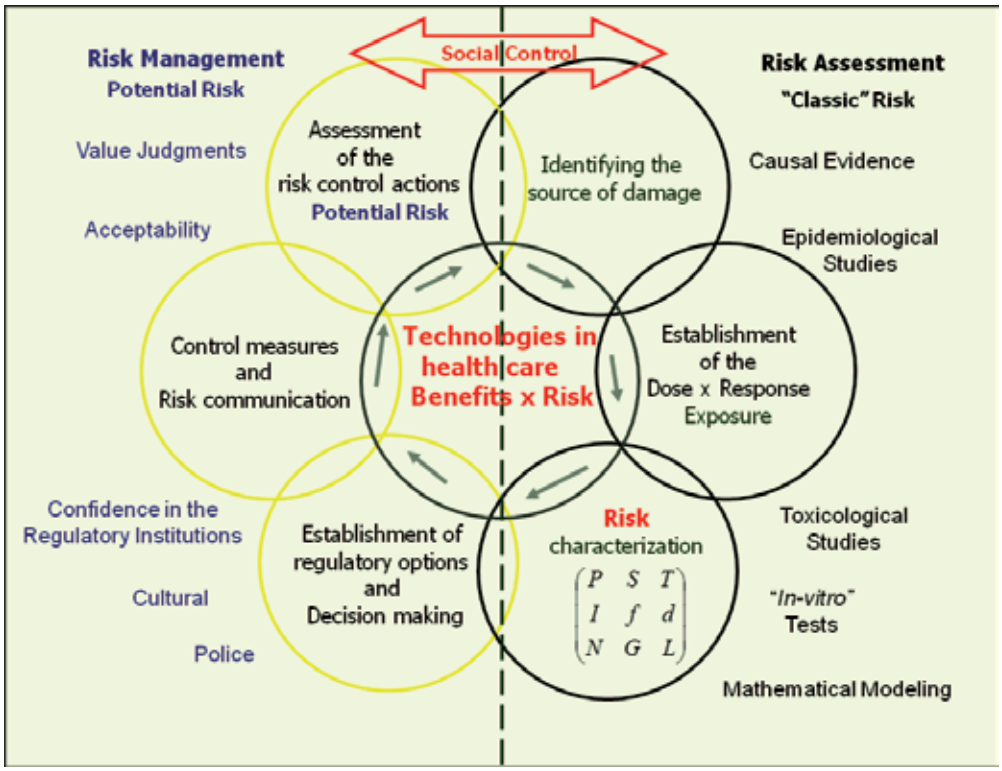


Fig. 1. Diagram of the paradigm of risks applied to the area of health surveillance. Adapted Omenn and Faustman (2005, p. 1084)

In the center of the map is the information that characterizes the particularization of the model for the health surveillance: the object of study. Objects of action of health surveillance, herein referred to as technologies in health care, have three basic characteristics: they are of interest to health, produce benefits and have intrinsic risks. It is these characteristics that justify the action of health surveillance about the technologies for health.

In this triad, the risk is a feature that mobilizes a wide set of control strategies. As the risk is intrinsic to the object, it cannot be eliminated without eliminating the object, it can only be minimized. All technologies for health present some kind of risk and, if there is any that does not possess risks, it probably will not be object of action of the sanitary surveillance.

For possessing risks inherent in their nature, the technologies should be used in the observance of the bioethical principle of the benefit (Costa, 2003, 2004)

The diagram of the paradigm of risk, represented in Figure 1, is divided in half, pierced by social control and the object of study. The right side represents the field of risk assessment and the left side, the field of risk management. Risk assessment is the use of objective evidences to define the effects on health due to exposure of individuals or populations to hazardous materials or situations. Risk management refers to the process of integrating the results of risk assessment with social, economical and political issues, weighing the alternatives and selecting the most appropriate to the regulatory action (National Research Council, 1983).

Risk assessment consists of three steps: identifying the source of damage, establishment of the dose x response and risk characterization. Risk identification is basically the answer to the question: which component of this health technology causes an adverse event? It is a question that can be answered based on causal, toxicological, and epidemiological evidence or in vitro tests (National Research Council, 1983, Omenn; Faustman, 2005).

In the second stage, two questions must be answered: how exposures occur? How is the relationship between exposure x effects (dose x response)? At this point, should be evaluated the conditions (intensity, frequency, duration, susceptibility and exposure period), in which the individuals or the populations are exposed. The second question should be answered with epidemiological, toxicological, experimental, and in vitro studies, using extrapolations or mathematical modeling, to establish the probability of occurrence (National Research Council, 1983, Omenn; Faustman, 2005).

The last step is the characterization of the risk, in the classic sense. It is a moment of synthesis, when setting the damage likely to occur and its probability (P) the severity of the damage (D), the lifetime lost (T) and the vulnerabilities of exposure, as the intensity of exposure (I), the frequency of exposure (F), the duration of exposure (D), the exposed population (N), the populational groups (G) and the accessibility to the geographical location of the population (L).

The risk assessment is a moment eminently technical and scientific, in which the theoretical models, the experimental procedures and the validation of the results are the elements of the performed studies (epidemiological, toxicological, in vitro and mathematical modeling, among others), so they can have rigor and scientific legitimacy. However, the evaluation models are not independent of the observers and their objectives (Czeresnia, 2004).

Risk assessment is not always possible to be performed quantitatively. In the case of the ionizing radiations, for example, the studied populations (Hiroshima and Nagasaki, Chernobyl and radiotherapy patients) were exposed to high doses, with high dose rates. Thus, it was necessary the use of the precautionary principle to postulate that, by extrapolation of the results of exposure at high doses, one must consider the linear relationship dose x response, without a threshold of exposure. Similar situations also occur in exposures to other physical and chemical elements, reflecting the complexity of the processes of risk assessment.

Based on information from the risk assessment, begins the process of management, conducted by the regulatory authority, also composed of three steps: establishment of regulatory options and decision making; implementation of control measures and risk communication and; assessment of the control actions.

In the first stage, are raised the possible actions that can minimize the risks, when the political-economical-cultural viability of each of the actions should be evaluated. Generally, there are several possibilities of regulation, when the best should be chosen. The best option is not, necessarily, the one with lowest risk or the one you want, it's the possible option in the evaluated context. The result of the value judgments will be the establishment of the limits of acceptability and of the control activities needed to keep the risks within these limits (National Research Council, 1983, Omenn; Faustman, 2005). In the case of the sanitary surveillance, this is the moment of development and publication of the standards for sanitary regulation.

The next step is the moment to inform society about the risks being regulated and the control measures being implemented. Parallel to the communication process, the regulatory authority should take the necessary measures, so that the control measures are effectively fulfilled by the regulated segment. An autonomous regulatory authority, with financial resources and skilled technicians, is a sine qua non condition for the implementation of the regulatory actions. However, the tradition of the institutions, of the regulated segment and of the society is essential so that risk control actions cease to be just rules and start to be practiced (National Research Council, 1983, Omenn; Faustman, 2005).

The last step is the evaluation of the entire process. It's the end of the first cycle and, perhaps, demands the beginning of a new cycle of risk assessment and management. To carry out the assessment, understood as a trial on a social practice or any of its components, in order to assist in decision-making, it is necessary to formulate strategies, select approaches, criteria, indicators and standards (Vieira Da Silva, 2005).

5. The potential risk

As seen so far, risk is a theoretical construct, historically grounded and, by the characteristics with which it presents itself in modern times, requires a regulatory system focused on protecting the health, due to the attributes that present the new technologies.

In the presented model of regulation of risks, the risk, in the classical sense, no longer has the central role, when passing from evaluation to management. In the process of risk management, the actions of health surveillance are focused, in general, on the control of risks and on the source of risks. In risk evaluation, the hazard is identified, related to the

damages and its consequences, thus risk is characterized. In risk management, the forms of control are identified, implemented and evaluated; thus control is characterized.

The sanitary standards generally do not regulate the action of chemical, physical or biological substances, they regulate actions, procedures, products and equipments that must be used, so that the technologies for health may produce the maximum of benefit with the minimum of risk, considering the scientific, ethical, economical, political and social issues.

The control actions are not related, necessarily, to the sources of risks. They may be related to conditions of the environment, of procedures, of human resources or of management of the own system of risk management. Since actions of health surveillance are focused, generally, on the control of risks and not on the risks itself, it becomes difficult the establishment of the cause-effect relationship.

The sanitary license, for example, is an operating concept that instrumentate the sanitary surveillance to control risk, but that is not directly related to any source of risk. A health service working without a sanitary license poses a risk to the system control, but may not represent a risk in the classical sense. One can not say what are the damages that may occur and in which probability. Even because the service can be fulfilling all technical and safety requirements. However, the absence of the license represents an unacceptable potential risk situation for the system control. Similar reasoning can be used to evaluate the equipment registration, the professional certification, among others.

The luminosity of the view box, used to view radiographic images, is another good example. The inadequate luminosity of the view box, despite not causing any direct harm to the patient, can hide radiological information and cause a misdiagnosis. In order to display the different tones of gray, in a radiography with optical density between 0.5 and 2.2, you need a view box with luminance between 2000 and 4000 nit[2]. So, what is the risk of using a view box with a luminance of 500 nit?

There are so many variables involved that the question becomes difficult to answer. The possibility of error or loss of diagnostic information, for example, cannot be understood as a harm to the patient. The damage will be done when the decision making of the medical procedure, based on incorrect or incomplete diagnostic information, is made effective. Thus, one cannot determine the damage that will be caused and what are the probabilities of occurrence. One cannot say, even, that damage will occur. However, it is an unacceptable potentially hazardous situation, as is known to the minimum necessary light in a view box, to produce a reliable diagnosis condition.

The potential risk concerns the possibility of an injury to health, without necessarily describing the injury and its probability of occurrence. It is an concept that expresses a value judgment about a potential exposure to a possible risk. It is as if it represents the risk of the risk.

It is observed that the potential risk passes to present itself as a possibility of occurrence, or an expectation of the unexpected, therefore, it's related with possibility and not with probability. This difference is crucial to be able to clarify the proposed concept, after all, the probable is a category of the possible, that is, something is only probable if it's possible,

[2] The unit of luminance in the International System is the cd/m2, known as nit.

because if it's impossible, you cannot talk about probable or improbable. This condition of potential risk demonstrates its anteriority in relation to the classic risk. In the examples above, one can not calculate the probability of a damaging event for the lack of sanitary license or the low luminosity of the negatoscope, but, given what is known, there are chances that harmful events may occur due to these conditions.

Another important feature of the concept of potential risk refers to the temporal dimension of causal relationships. While the classic risk has its evaluation basis in occurred events, the potential risk has its causal evaluation foundations in the events that are occurring and the effects that may, or may not, occur in the future. Thus, allows work with the temporal dimension of risk facing the future or for a meta-reality and not for the past.

It is also possible to differentiate the potential risk from the classical risk according to the strategies used in the public health practices. These strategies can be divided into three great groups: health promotion in the restricted sense, health prevention (of risks or damages) and health protection.

In the practices of health promotion, strategies are aimed at capacity building and at raising awareness of the groups, so that they can take action to improve the quality of life and health, without being directed to a disease or injury whatsoever. They are actions of an educational nature which are not related to one or another specific risk factor (Almeida Filho, 2008). Thus, as their strategies do not involve specific risk factors, remains to discuss the concept of risk involving the two other strategies.

Regarding the preventive health strategy, the search for the determinants or the risk factors of a disease or of a specific aggravation on temporally and spatially defined individuals characterize their actions.In other words, are destined to act on these factors in order to reduce or eliminate new occurrences in the collective.It starts from "the assumption of recurrence of events in series, implying in an expectation of stability of the patterns of serial occurrence of the epidemiological facts" (Almeida Filho, 2000). As the action is given according to specific risk factors, ie, is related to the known behavior of the cause (risk factor) according to the probability of occurrence of the unwanted effect, the classical concept of risk seems to be the most appropriate.

On the other hand, health protection is intended to strengthen the individual defenses, therefore, is not always directed to known causes and specific risks, or relate to the referred events in series. They are used, in most cases, when there is an epistemic uncertainty, ie, when it's unknown or there is little information about the problem to be resolved or a decision to make. So, in the case of the health protection strategies, the central element in risk management is the potential risk that, despite not, necessarily, representing a defined relationship of cause and effect, can be quantified and classified into levels of acceptability, as will be discussed further, becoming an important operational concept of the sanitary surveillance.

However, the potential risk, as well as the classic risk, cannot be represented in most scientific fields by only a number. It should be understood and evaluated within a context and with limits of acceptability established by the technical and social determinants. Therefore, the evaluations made by regulatory authorities in the process of risk management have as indicators, in most cases, the tools of risk control and, as consequence, a measure of potential risk, which will indicate whether the control conditions are acceptable or not.

6. Strategy for operationalization of potential risk

The operationalization of the concept of potential risk has implications for the sanitary surveillance, because the quantification, classification and definition of acceptability levels of these risks will permit the monitoring and comparison of several objects under the control of the sanitary surveillance, such as, the health services.

A strategy to operationalize this concept is to establish a mathematical function that relates potential risk with risk control indicators. These control indicators are present in the rules, ie, are the characteristics associated with equipments, procedures, health services etc., that should be controlled within the pre-established parameters.

The control indicators represent elements that, in most cases, you do not know the probability of generation of harmful effects, but, if outside of the pre-established parameters, there is a possibility that a harmful event may occur. Therefore, there is a causal relationship between indicators of control and potential risk, where both are inversely proportional, ie, the closer to the predetermined values are the control indicator, the lower the potential risk and vice versa.

Having identified the causal relationship it's possible to establish mathematical formulations that describe the behavior of these relationships, through the traditional mathematical formalism or using new theoretical contributions to the theory of fuzzy sets which together with the theories of evidence and of possibility, constitute a new field of study that aims at the treatment of epistemic uncertainties within the possibilities, as will be shown below.

6.1 A fuzzy logic system to evaluate potential risk

The theory of fuzzy sets, developed by Zadeh (1965), was born from the observation that in the real world certain objects or beings, such as the bacteria, are ambiguous as to which class they belong to, ie, have characteristics of animals and also vegetables. The observation of this ambiguity has led to the thought that there is no precision in the limits of a set and thus, it is possible to establish degrees of belonging of an element X, whatever, to a certain set. Taking as an example the bacteria, the number of animals characteristics that they exhibit allows us to establish a degree of belonging to the set of the animals, as well as, the amount of plant characteristics allows us to establish another degree of belonging to the set of the vegetables.This way, although they have a higher number of features of one kind or another, the bacterium does not cease to belong to both, though with different degrees of belonging.

However, in the analysis of the ambiguities present in most of the everyday phenomena, is not always possible to quantify the characteristics of an element with precision to determine its degree of belonging. In most cases, these characteristics are presented in the form of uncertainties. To solve this problem, the modeling of the uncertainties uses the natural language (ordinary) and the membership functions express the possible values between 0 and 1, which each natural term may take. (Weber,2003).

As in natural language are used variables or linguistic terms, also called inaccurate quantifiers , of common use in everyday life, but definers of many decisions, such as, "low," "high," "good," "very good"," tolerable "and so on. The membership functions consist of the association of each linguistic variable to a standard curve of possibilities (Shaw; Simões, 1999), which will define the membership degrees between 0 and 1, that the linguistic variable may assume.

Zadeh (1965) developed operators for the *fuzzy* sets, enabling the establishment of relationships between them, being the most important the operations of maximum (max) and minimum (min), which can be easily understood if defined, respectively, as union and intersection in the classical set theory.

A fuzzy logic system (FLS), in a simplified manner, consists of performing logical operations with several fuzzy linguistic variables, in order to obtain a single value that represents the result of the performed operations.

To build an FLS, the first step consists of the definition of the input and output variables of the FLS, depending on the problem you want solved. When you want to, for example, know what is the potential risk indicator of biological contamination of the water for dialysis in the realization of the hemodialysis procedure; the output variable of the FLS may already be defined as the potential risk indicator of biological contamination of the water for dialysis (PRI-BCW)

To establish the input variables, the first question to be answered is: what are the possible causes to make water for dialysis potentially dangerous for biological contamination? Loosely, we can say that there are four causes: 1) Inadequacy of the drinking water treatment; 2) Inadequacy of the water treatment for dialysis; 3) Lack of knowledge or error of an employee who performs the procedure of water treatment for dialysis and 4) Inadequacy on the facilities of the water treatment plant.

The second question, in an attempt to define the input variables, is: how to handle each cause defined? Consulting the existing regulations for dialysis services in Brazil, you can display at least one control point for each defined cause, as described in Table 1.

Cause	Control point
Inadequacy of the drinking water treatment.	Adequacy of the procedure for drinking water treatment, according to the Ordinance MS nº 518/2004 4
Inadequacy of the water treatment for dialysis.	Adequacy of the execution of the procedure of water treatment for dialysis, according to the RDC nº 154/2004 5
Lack of knowledge or error of an employee who performs the procedure of water treatment for dialysis.	Adequacy of the capacity of an employee who performs the procedure of water treatment for dialysis.
Inadequacy on the facilities of the water treatment plant.	Adequacy of the constructive aspects and of the equipment used in the water treatment plant, according to the RDC nº 154/2004 5

Table 1. Relationship between possible causes and control points of the possibility of biological contamination of the water for dialysis.

Established the control points of the four possible causes, it is up to define which input variables of the FLS will be the results of the verification of the level of control of the set points. This level of control is called control indicator (CI) and shall be established by an observer, such as, a public health professional with expertise to make a subjective evaluation of each item, and may be defined, therefore, for a *fuzzy* linguistic variable, or inaccurate quantifier.

Defined the input and output variables of the FLS, it is necessary to establish the universe of discourse of each of them, ie, the variation range of the fuzzy linguistic variables of input and output. The universe of discourse limits the possible evaluations that the observer can present. As is the case of the input variables of the FLS, it is to check its adequacy, we will use the universe of discourse in terms of: Inadequate (IND), Shortly Adequate (SAD), Tolerable (TOL), Adequate (ADQ) and Very Adequate (VAD).

For the output variable of the FLS, since it is an indicator of potential risk, the universe of discourse adopted will be: Very Low (VL), Low (L), Medium (M), High (H) and Very High (VH). Note that in all cases the universe of discourse consists of 5 variables to allow good accuracy, since the greater the number of possibilities is, the better the accuracy of the evaluator.

The next step will be to define the logical operations that must be made in the FLS so that, from the input variables, it can be obtained the potential risk indicator of biological contamination of the water for dialysis (PRI-BCW) in the output. Being the four input variables of the type, verification of the "level of adequacy", and as the output variable should represent an indicator of potential risk, two questions must be evaluated: which operations should be performed between the four input variables? and what is the relationship between control indicator (CI) and potential risk indicator (PRI)?

The operation between the input variables of the FLS should be held so that it is possible to obtain a single value, ie, a value that represents the level of control of all input variables (control indicator), ie, an indicator of aggregate control. Therefore, this must be one of the logical operations to be performed.

The control indicators represent the level of control found by the observer and the 'potential risk' is the output of the FLS. Thus, the indicator of potential risk is inversely proportional to the control indicator, since the greater the observed control indicator, the lower the potential risk and vice versa. So, this will be another operation to perform

To perform these operations, will be used *fuzzy* logic controllers. A *fuzzy* logic controller is a device that performs logical operations between *fuzzy* linguistic variables in its three stages: fuzzification, *fuzzy* inference and defuzzification. In this case, you need to build two types of *fuzzy* logic controllers, one for each type of operation you need to perform.

For each of the *fuzzy* controllers, it is necessary to develop the three steps referred above (fuzzification, *fuzzy* inference and defuzzification); therefore, it will be demonstrated, initially, the operation between the input variables of the FLS, known only as input controller. Each controller must perform only the operation between two input variables, so there is no explosion of rules, as will be explained later.

Fuzzification means the process of transforming the possible existing information into *fuzzy* elements; consists in identifying the linguistic variables of input and output that you want to operate, defining the universe of discourse and the membership functions for each variable, based on the experience and on the nature of the process being fuzzified.

To perform the fuzzification of the input controller, some steps have been taken, as the identification of the input linguistic variables and the establishment of the universe of discourse (Inadequate – IND, Shortly Adequate – SAD, Tolerable – TOL, Adequate – ADQ and Very Adequate – VAD). However, it lacks defining the output variable and the universe of discourse for this controller, because, as has been identified, it will be required more than one logical operation between the *fuzzy* variables, the output of this input controller, will not, necessarily, be equal to the output variable of the FLS. Thus, considering that the objective of this controller is to aggregate the control indicators (CI) pointed by the observer and thinking about the future composition of the organization of the FLS, it was decided, in the example shown, to define the universe of discourse of the output variable as: Very Low (VL), Low (L), Medium (M), High (H) and Very High (VH).

The last step to accomplish the process of fuzzification is to define the membership function for each identified *fuzzy* linguistic variable. In this case, we took the function of trapezoidal and symmetrical shape for all the input controller's *fuzzy* linguistic variables, as can be seen in Figure 2.

A membership function defines the degree of belonging or membership of each *fuzzy* linguistic value, ie, it represents the curve of possibilities of the behavior of the *fuzzy* linguistic variable (Weber, 2003). Note that the membership functions are standard functions, ie, in its ordinate axis (Y) it only admits *fuzzy* values from '0 'to '1', ie, it goes from the not belonging (0%) to the total belonging (100%). In the abscissa axis (X) the values depend on the problem addressed; in this case, we used '0 'to '1', because those are variables that assume this behavior (potential risk and control indicator).

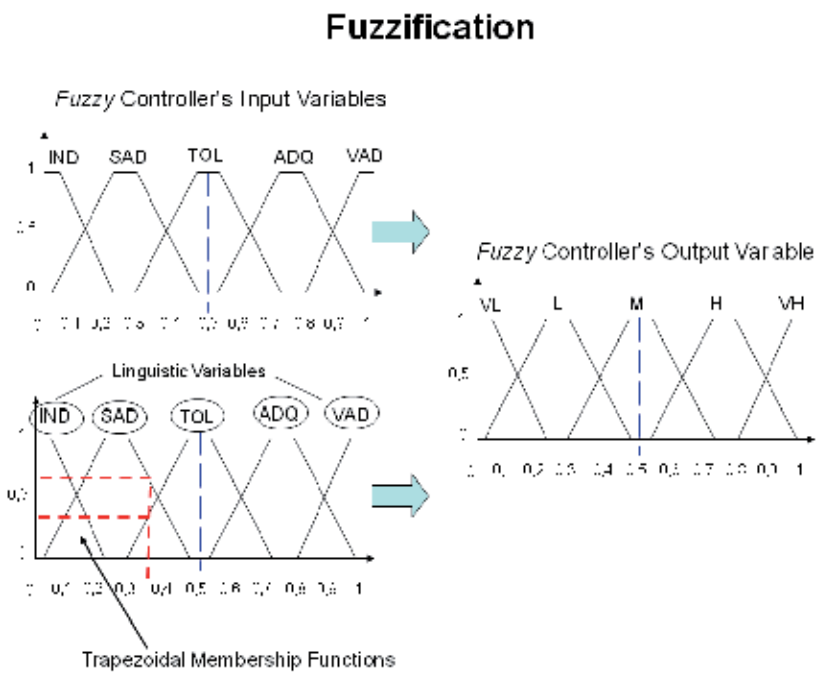


Fig. 2. Input controller's input and output membership functions

It is also important to point out, in Figure 2, that each *fuzzy* linguistic variable was associated with a numerical value 0%; 25%, 50%, 75% and 100%, respectively. This fact can be identified, by observing that the top of the trapezoids corresponds to one of these values. So, if the point 0.5 is taken (50%), in the X-axis (blue dotted line), it will correspond to the center of the trapezoid for the *fuzzy* linguistic variables 'Tolerable', in the input and 'Medium', in the output.

We opted for the trapezoidal shape because it was recognized that in the observation made there is no accuracy of values; when reporting, for example, that a level of control is 'shortly adequate', this does not correspond exactly to 25% but to a range for that value. Now the option for the symmetry was made as it was considered that there are an equal number of chances of the observer to choose for any of the *fuzzy* linguistic variables that compose the universe of discourse.

The importance of fuzzification can be understood, when we take a value, for example, 0.35 in the abscissa axis (red line), note that this value has a degree of membership greater than 50% for 'shortly adequate' and less than 50% for 'tolerable'. These membership differences will generate the sets that will be operationalized.

Completed the process of fuzzification, it is necessary to perform the *fuzzy* inference process. The *fuzzy* inference process consists in the processing of the *fuzzy* variables according to specific rules. There are basically two methods, the Mamdani model and the Takagi-Sugeno-Kang model (Shaw; Simões, 2005). Mamdani's method is the most used and recommended for the treatment with inaccurate information. It is based on the elaboration of rules of the 'IF' <condition>; 'THEN' <consequence> type, using the heuristic method. The rules are the knowledge bases, from which, an "inference machine" (*software* or *hardware*) acts and performs operations of minimum (intersection) between the input *fuzzy* linguistic variables of each rule, and of maximum (union) between the results obtained by the previous operation.

A rule of the 'IF' <condition>; 'THEN' <consequence> type is a simple logic rule and it means that for a given situation, 'IF' a condition is met, even partially, 'THEN', a consequence will occur. For example, when one states that the potential risk is inversely proportional to the level of control, it is possible to say that 'IF' the level of control is high, 'THEN' the potential risk is low. When two variables (two conditions) are associated, using the Mamdani method, we use the operator 'AND' between the two variables to indicate that an operation will take place between them. This way, the rule is now stated as: 'IF' a condition is met, even partially 'AND' other condition is also met, even partially, 'THEN', some consequence will occur. This way, using the heuristic method was constructed the rules base for *fuzzy* logic controller input, shown in Table 2.

It will be required the construction of twenty-five rules, because for two variables per controller and five fuzzy linguistic variables, one needs, therefore, twenty-five combinations (5^2).

It should be noted, also, that in Table 2 the input variables were treated generically as 'Adequacy 1' and 'Adequacy 2', because it will be necessary to use more than one input controller, since there are four input variables. So, you can use the same set of rules for both controllers.

The 'inference machine' is a *software* or *hardware* that performs logic operations based on defined rules.

Rules	IF	Condition	AND	Condition	THEN	Condition
1	'Adequacy 1'	VAD	'Adequacy 2'	VAD	'Control'	VH
2	'Adequacy 1'	VAD	'Adequacy 2'	ADQ	'Control'	VH
3	'Adequacy 1'	VAD	'Adequacy 2'	TOL	'Control'	M
4	'Adequacy 1'	VAD	'Adequacy 2'	SAD	'Control'	L
5	'Adequacy 1'	VAD	'Adequacy 2'	IND	'Control'	L
6	'Adequacy 1'	ADQ	'Adequacy 2'	VAD	'Control'	VH
7	'Adequacy 1'	ADQ	'Adequacy 2'	ADQ	'Control'	H
8	'Adequacy 1'	ADQ	'Adequacy 2'	TOL	'Control'	M
9	'Adequacy 1'	ADQ	'Adequacy 2'	SAD	'Control'	L
10	'Adequacy 1'	ADQ	'Adequacy 2'	IND	'Control'	L
11	'Adequacy 1'	TOL	'Adequacy 2'	VAD	'Control'	M
12	'Adequacy 1'	TOL	'Adequacy 2'	ADQ	'Control'	M
13	'Adequacy 1'	TOL	'Adequacy 2'	TOL	'Control'	M
14	'Adequacy 1'	TOL	'Adequacy 2'	SAD	'Control'	L
15	'Adequacy 1'	TOL	'Adequacy 2'	IND	'Control'	VL
16	'Adequacy 1'	SAD	'Adequacy 2'	VAD	'Control'	L
17	'Adequacy 1'	SAD	'Adequacy 2'	ADQ	'Control'	L
18	'Adequacy 1'	SAD	'Adequacy 2'	TOL	'Control'	L
19	'Adequacy 1'	SAD	'Adequacy 2'	SAD	'Control'	VL
20	'Adequacy 1'	SAD	'Adequacy 2'	IND	'Control'	VL
21	'Adequacy 1'	IND	'Adequacy 2'	VAD	'Control'	L
22	'Adequacy 1'	IND	'Adequacy 2'	ADQ	'Control'	L
23	'Adequacy 1'	IND	'Adequacy 2'	TOL	'Control'	VL
24	'Adequacy 1'	IND	'Adequacy 2'	SAD	'Control'	VL
25	'Adequacy 1'	IND	'Adequacy 2'	IND	'Control'	VL

(Inadequate (IND), Shortly Adequate (SAD), Tolerable (TOL), Adequate (ADQ), Very Adequate (VAD), Very Low (VL), Low (L), Medium (M), High (H) and Very High (VH)).

Table 2. Rules 'IF'...'THEN' for *fuzzy* input controller

As shown in the example above, when it was shown the importance of fuzzification, when defining a control indicator for an input variable, it will be associated with a number that will produce different degrees of membership for each membership function and, at every point where it intercepts the membership function, it will generate *fuzzy* sets. In the *fuzzy* inference it is verified if there is a point of interception for all defined rules 'IF', 'THEN'. Among the sets generated in each variable and in each rule, it is performed an operation of minimum (intersection) that corresponds to the operator 'AND'. Among the resulting sets from the operation of minimum of every rule, it is performed an operation of maximum (union), coming to a set representing the results of the performed *fuzzy* operations.

In Figure 3, it is shown what happens in the process of *fuzzy* inference. It was assigned to the 'adequacy of the procedure for drinking water treatment' (ADWT) a control indicator 'Tolerable' (0.5) and to the 'adequacy of the procedure of water treatment for dialysis'

(AWTD), a control indicator 'Adequate' (0.75). The red lines represent the values assigned to each input variable and the yellow forms, the set generated in each rule. Note that operations of minimum (intersection) are performed between the yellow sets of each rule, generating as results the blue sets. Among the blue sets, an operation of maximum (union) is performed, resulting in the set surrounded by a red line, representing the *fuzzy* result.

The defuzzification process is translated into the transformation of the *fuzzy* set resulting in a discrete value, seeking to define the value that best represents the distribution of possibilities present in the output variable. The three most used methods for defuzzification are the center of area (C-O-A), the center of maximum (C-O-M) and the mean of maximum (M-O-M). The C-O-A method calculates the centroid of the area obtained in the output, or the point that divides this area in half, after the max-min operations performed on *fuzzy* inference. The C-O-M method calculates a weighted average of the maximum values present in the exit area, which weights are the results of *fuzzy* inference, the area itself has no influence on the outcome. Finally, the M-O-M method, used in this work, calculates an average of the maximum values present in the exit area, disregarding the format of this area, as shown in Figure 3.

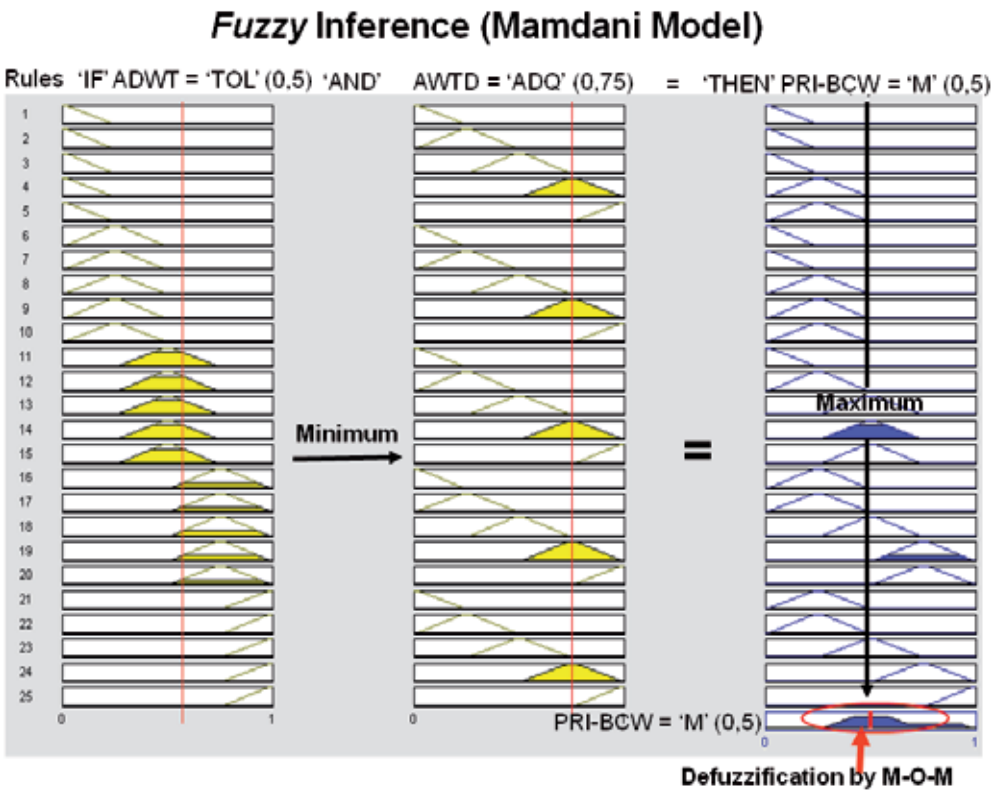
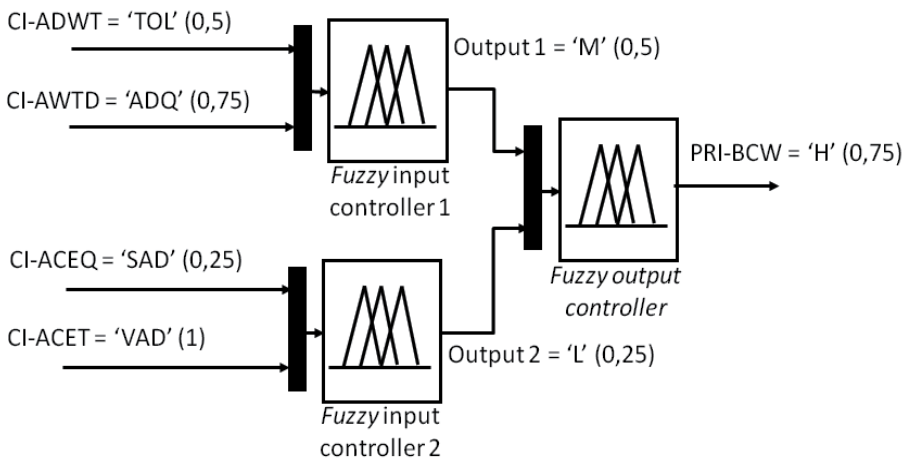


Fig. 3. The steps of fuzzy inference and defuzzification for the input controller

The second type of *fuzzy* logic controller to be built is called output controller. As can be seen in Figure 4, the input variables of this output controller will be equal to the output variables of the input controller and the output variables will be equal to the output variables of the FLS. The only difference will be the rule base 'IF'; 'THEN', but all other steps are identical to the input controller. The difference in the rule base exists, because the logical operation to be performed will be the conversion of the indicator of control for potential risk indicators that are inversely proportional. Thus, the rule base 'IF', 'THEN' was elaborated considering this criterion.

Finally, for the construction of the FLS the *fuzzy* logic controllers will be grouped so as to produce the desired information, as shown in Figure 4. To carry out the construction and operation of an FLS, the program MatLab can be used.



CI-ADWT = Control indicator of the adequacy of the procedure for drinking water treatment

CI-AWTD = Control indicator of the adequacy of the procedure of water treatment for dialysis

CI-ACEQ = Control indicator of the adequacy of the constructive aspects and of the equipment used in the water treatment plant

CI-ACET = Control indicator of the adequacy of the capacity of an employee who performs the procedure of water treatment for dialysis

PRI-BCW = Potential risk indicator of biological contamination of the water for dialysis

Fig. 4. Fuzzy logic system for indication of potential risk of biological contamination of the water for dialysis

Thus, as can be seen in Figure 4, when evaluating a service of dialysis a sanitary inspection team should consider the control indicators of the adequacy of the procedure for drinking water treatment (CI-ADWT), of the adequacy of the procedure of water treatment for

dialysis (CI-AWTD), of the adequacy of the constructive aspects and of the equipment used in the water treatment plant (CI-ACEQ) and of the adequacy of the capacity of an employee who performs the procedure of water treatment for dialysis (CI-ACET), respectively, 'TOL', 'ADQ', 'SAD' e 'VAD', so, the PRI-BCW of this system will be considered high (H), ie, 0.75; indicating that there is a nonconformity at some point in the process under analysis. In this case, the inadequacy of the constructive aspects of the water treatment plant and / or of the equipment used to perform the process.

6.2 O PRAM: Potential Risk Assessment Model

The formulation of the PRAM has been developed generalized so that it could be applied in any area of risk governance and possibly, also outside it.

The PRAM was validated by evaluating potential risks in radiodiagnostic services in the State of Bahia, Brazil, enabling advance, in order to better understand the specific problems and the possibilities of action of the health surveillance system, as the regulatory authority, in control of risks in radiodiagnostic.

The validation results showed that use of the PRAM model allowed going beyond simple situational description, indicating the possible explanatory factors of the health situation found. Some advantages of this approach are introduced, in comparison with other works that dealt with the theme. One of them concerns the graphical representation of the potential risk of each procedure in each of the services.

This enables the regulatory system to classify and compare the evaluated procedures, so that you can plan and direct the actions for the services whose procedures are in unacceptable or tolerable level of potential risk , establishing priorities.

Another advantage relates to the possibility of applying the principle of optimization in the risk control system, enabling the continuous evolution of the system, evaluating the historical evolution of risk management. The monitoring of time evolution can show an advance or a retreat of the potential hazard, alerting the regulatory authority before the service moves to a range of higher degree of risk, allowing risks prevention actions, by anticipating and stopping a trend.

So, the Regulatory Authority has the possibility to act in preventing the risk and not just in control. The temporal evolution can be used easily, with computational aid, to monitor the services individually or collectively.

However, using PRAM to monitor the temporal evolution of the potential risks, as well as for comparison and risk assessment, should be carried out using the same rating scales and indicators of the same ranges of acceptability. Otherwise, the PRAM loses comparability.

The PRAM needs to consider important issues of risk governance. The first question refers to the range of variation. The PRAM needs to be represented by a mathematical formalism, whose values of the potential risk - PR are always within the same range of variation, regardless of the number of indicators, and there is no possibility of taking the zero value.

The issue of the values being within the same range of variation allows the comparison and the establishment of limits of acceptability, while the not possibility of assuming the value zero is a condition of the problem, because the risks can be as small as possible, but will never be nulls.

The levels of acceptability should not have a direct border between the acceptable and the unacceptable. There should be a transition zone, where the condition of risk is tolerable in certain conditions or for some time. The levels of acceptability must permit its variation, for more or for less, allowing the application of the principle of optimization (Slovic, 2000).

On the other hand, the number of indicators should be opened, allowing the inclusion and exclusion of as many indicators as may be necessary. The indicators are classified, according to the level of potential risk they pose to the system.

The risk control indicators should be separated into two categories: critical indicators and non-critical indicators. Critical indicators are those that are associated, directly, to the unacceptable potential risk level. For its severity, they compromise the whole risk control system of the procedures. Therefore, report about critical situations, whose existence, regardless of the existence of any other, take the potential risk to the unacceptable levels.

The set of the non-critical indicators is formed by all the indicators that, individually, do not compromise, in a decisive way, the risk control system. The complete set of the non-critical indicators acts like a critical indicator, ie, if all non-critical indicators are null, the set of indicators will be null and thus, only then, will represent a critical commitment on the potential risks control system.

Once one can build as many risk indicators as needed or desired and the result must be within fixed limits, fundamental to the discussion and establishment of acceptability criteria of the potential risks, it was necessary to develop a mathematical formalism to represent the mean values of the sets of indicators (critical and noncritical) through a single value.

The set of critical indicators is formed by I_C the indicators

$$\left\{C_{I_1}\,;\,C_{I_2}\,;\,C_{I3}\,;...;C_{IN}\right\} \tag{1}$$

Since the critical indicators have the ability to compromise the entire potential risk control of the system, as well as they need to be represented by a mean, the most appropriate way is to represent them as a geometric mean. The geometric mean is the nth root of the product of N terms, representing a mean value of the product. Thus, to represent a mean of N terms, we have:

$$\overline{C}_I = \sqrt[N]{\prod_{i=1}^{N} C_{Ii}} \tag{2}$$

So, if any of the indicators has zero value, the value of I_C will be zero, independent of the other indicators. On the other hand, the maximum value is, numerically, equal to the

maximum value of an indicator, ie, regardless of the number of indicators that is selected, the result will always be in the same range of variation.

The set of non-critical indicators is formed by the I_{NC} indicators.

$$\left\{NC_{I_1}\,;\,NC_{I_2}\,;\,NC_{I_3}\,;...;NC_N\right\} \tag{3}$$

Once the non-critical indicators do not have the ability to, individually, represent the commitment of all the system potential risks control, cannot have its mean represented by a multiplicand. However, they also need to be represented by a mean, so that the representative value of the set is equal, at most, to the maximum value of one of its elements and is within a known range of variation.

Therefore, the best way to represent them is through an arithmetic mean. The non-critical indicators (INC) can be represented by a simple arithmetic average, because it can only be zero, if all control indicators are non-existent.

$$\bar{N}C_I = \frac{\sum_{j=1}^{M} NC_j}{M} \tag{4}$$

The function risk control (RC), which represents the result of the indicators of risks control, should be represented as the geometric mean, ie:

$$R_C\left(C_I, NC_I\right) = \sqrt{\bar{C} \times NC} \tag{5}$$

Once more, we used the geometric mean, so that the risk control (RC) is in a range of variation known in advance and that depends only on the variation of I_C and I_{NC}.

Taking the risk control (RC) as the independent variable, the function that best represent the relationship of cause and effect between risk control and potential risk is the exponential function, with the following form:

$$P_R(R_C) = e^{-R_c} \tag{6}$$

PR (RC) - Potential risk function, which is dependent on the risk control function, will be referred to as PR; RC - Risk control, function that determines the potential risk and that, on the other hand, is determined by the indicators of risk control.

The shape of the exponential function, with a rapid decrease, represents a good model for critical phenomena, as is the case of the potential risk for health services. The complex relationship between the various factors that influence in the risk control exhibits a kind of not extensive sum, where the potential risk for an event, involving the junction between two factors, can be greater than the sum of the potential risk of the two factors separately.

This type of behavior ends up generating a sudden increase of the potential risk, when adding many elements or some critics, being perfectly represented by the rapid decrease of the exponential function.

Another important behavior of the exponential function, to represent the potential risk, is that it has a finite maximum value and the minimum value tends to zero, without necessarily assuming the zero value. The potential risk of a system cannot increase indefinitely, and cannot be zero. Its possibility of occurrence is finite and, for bigger and better that it is the risk control system, you cannot reach a situation of absence of potential risk.

The function proposed in this article, represented by equation (6), allows the potential risk to vary between the maximum value 1 and the minimum value that will be defined by the risk control indicator. The minimum value, will never be zero and, regardless of the number of indicators that it is used, the potential risk function will have fixed maximum and minimum values.

So, an important issue in this model is to establish the range of variation of the risk control indicators, as the maximum scale value defines the minimum value that the potential risk function (PR) can take and, consequently, its range of variation. It is worth noting that the potential risk assessments with this model can only be compared, if they use the same scale of variation of the risk control indicators.

The I_C and I_{NC} indicators are evaluated, on a scale of zero to five, where zero represents non-existent or inadequate risk control and five represents risk control excellent, with the following degrees: 0 – absent or inadequate; 1 – poorly; 2 – reasonable; 3 – good; 4 – great and 5 – excellent.

One should consider that the compliance with the rule is associated with the value 3. Thus, regardless of the number of critical and non-critical indicators, the risk control function (RC) will assume values, necessarily, between 0 and 5. Then, the maximum and minimum values of the potential risk (PR) will be:

$$P_R(R_C=0) = e^{-0} = 1{,}000 \quad (7)$$

$$P_R(R_C=5) = e^{-5} = 0{,}007 \quad (8)$$

When RC = 0, which means the absence of the set of non-critical risk controls or the absence of one of the critical risks controls, the potential risk will be PR (0) = 1, ie, there is a full potential risk situation. One can describe the possible potential damage; yet one can not specify a damage and its associated probability of occurrence. On the other hand, for greater that are the controls, the potential risk (PR) will never assume the zero value.

So, one can insert or remove as many risk control indicators as may be necessary, whether they are critical indicators or not, there will be no change in the variation of the function ($0.007 \leq PR \leq 1.000$).

The exponential function proves to be adequate to describe risk control systems, because it reflects well the concept of risks inherent to the technologies, ie, the risk can and should be minimized ever more, but can not be totally eliminated, because it is part of the technology itself. Ie, even if they have implemented all risk control mechanisms, it has a minimum potential risk value (intrinsic), which can not be eliminated, being that the benefits justify the use of this technology for health.

The RC function can also be understood as the relationship between the macro and micro indicators of the service. The means I_C and I_{NC} contain all the information service, so that they behave as if they were the micro systems states, that compose a given health service, determined by the individual indicators I_C and I_{NC}. Through them, we can know the situation of the equipment, of the human resources or of the procedures, while RC reports a macro value, aggregated, indicating the situation of the total risk control service, but nothing about its components, specifically. Both, RC and I_C or $I_{CN,}$ are of fundamental importance for the understanding of the risk control situation, depending on who is looking and what you want to analyze.

As the potential risk (PR) cannot be understood only as a dimensionless number more information are needed to support a decision making. As a way to aggregate the dimension acceptability, the potential risk should be represented within an area of potential risk with their respective bands of acceptability, as shown in Figure 5.

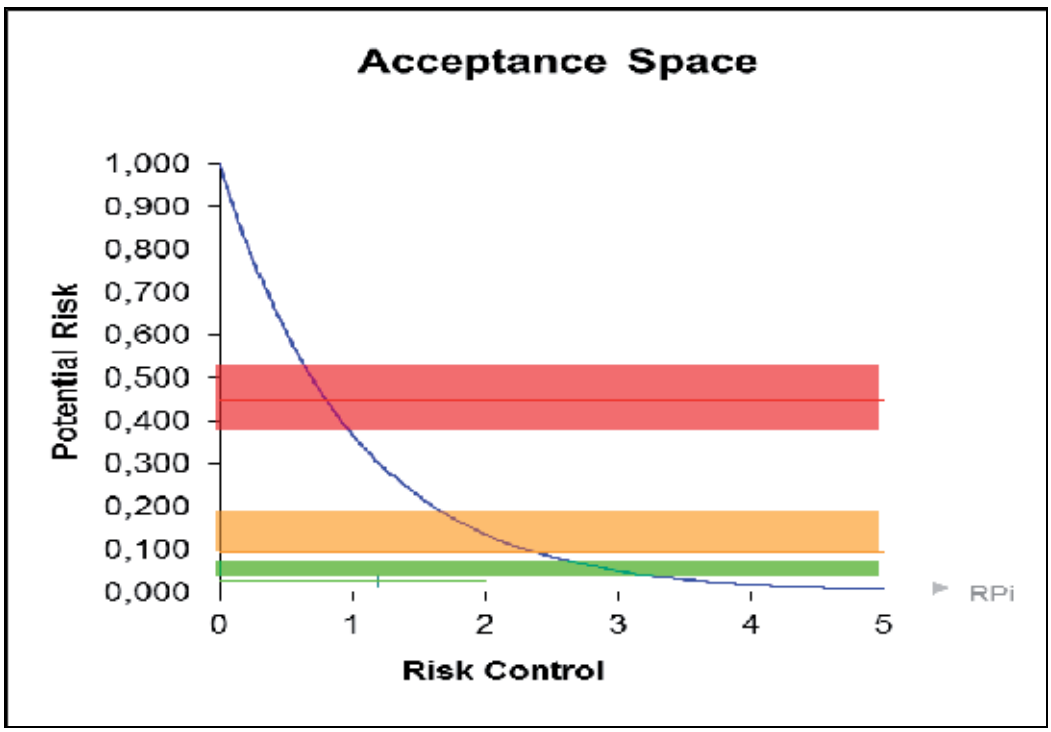


Fig. 5. Risk acceptance space of the PRAM

The idea of risks space was first proposed by Slovic et al. (1979), to perform a comparison of the perception of different types of risks and how experts and lay people perceive risks, by using psychometry to quantify the technologies, understood, in the broadest sense, such as equipment, products, processes or practices.

As there is a possibility of more than one evaluation with the same value of potential risk, causing a point overlap in the spatial representation, you can add a pie chart, so that you can see the number of services / procedures evaluated.

The IRGC "International Risk Governance Council" in the "white paper nº2", of 2006, proposes a bidimensional graphical representation to classify the risk levels of the nanotechnologies, using a non-linear representation, ranges of acceptability and a undefined region between the lower limit of the curve and the X-axis. It is a qualitative representation without estimation of values, which is meant to represent the shape of risks behavior in nanotechnology and its acceptability (IRGC, 2006). The work points to the need for quantitative graphical representation, which seems to have bumped in the difficulty to mathematically formulate the model. This difficulty was surpassed with the presented formulation of potential risk.

7. Conclusion

The concept of potential risk regards the possibility of occurrence of a health problem, without necessarily describing the injury and its probability of occurrence. It is a concept that expresses the value judgment about potential exposure to a possible risk. It's like representing the risk of the risk.

An important aspect of the concept of potential risk refers to the temporal dimension of causal relationships. While the classical risk has its basis of evaluation in occurred events, the potential risk has its causal bases of evaluation in the events that are occurring and in the effects that may, or may not, occur in the future. Thus, allows working with the temporal dimension of risk facing the future or a meta-reality and not the past.

In the case of the inspections of the health regulatory authorities, the central element in risk management should be the potential risk that, although not representing, necessarily, a defined relation of cause and effect, can be quantified and classified into levels of acceptability, as discussed in the presented model.

However, the potential risk, as the classical risk, can not be represented, only, by a number. It should be understood and evaluated within a context and with limits of acceptability established by the technical and social determinants. Therefore, the evaluations made by regulatory authorities in the process of risk management as indicators have, in most cases, the tools of risk control and as a consequence, a measure of potential risk, which will indicate whether the control conditions are acceptable or not.

8. Acknowledgments

This publication was financed by the Federal Institute of Education, Science and Technology of Bahia

This study is part of INCT-Citecs funded by the National Intitutes of Science and Technology Programme (MCT-CNPq, Brazil). Contract no. 57386/2008-9.

The authors thank especially Drs. Gunter Drexler and Ediná Alves Costa.

9. References

Almeida Filho, N. *A ciência da saúde*. São Paulo: Hucitec, 2000.

_____. O conceito de saúde e a vigilância sanitária: notas para a compreensão de um conjunto organizado de práticas de saúde. In: Costa, E. A. (Org.). *Vigilância sanitária: desvendando o enigma*. Salvador: EDUFBA, 2008. p. 19-43.

Beck, U. World risck Society. Cambridge: Polity Press, 2003.

Costa, E. A. Vigilância sanitária: proteção e defesa da saúde. In: Rouquayrol, M.Z.; Almeida Filho, N. *Epidemiologia & saúde*. São Paulo: Medsi, 2003. p. 357-87.

_____. *Vigilância sanitária: proteção e defesa da saúde.* São Paulo: Sobravime, 2004.

Covello, V. T., Munpower, J. Risk analysis and risk management: an historical perspective. *Risk Analysis*, v. 5, n. 2, 1985.

Czeresnia, D. Ciência, Técnica e Cultura: relações entre riscos e práticas de saúde. *Cadernos de Saúde Pública,* São Paulo, *v*. 20, n. 2, p. 447-455, 2004.

Fischhoff, B et al. *Acceptable risk*. Cambridge: Cambridge University Press, 1983.

Fischhoff, B; Bostrum, A; Quadrel, M. J.Risk perception and communication. In: Detels, Roger et al. *Oxford Textbook of Public Health.* 4 th. New York: Oxford University Press, 2005. v.1.

Gelman, A.; Nolan, D. *Teaching statistic a bag of tricks.* London: Oxford, 2004.

Hampel, J. Different concepts of risk: a challenge for risk commnication. *International Journal of Microbiology*, n. 296, p. 5-10, 2006.

Hood, C.; Rothstein H.; Baldwin, R. *The government of risk: understanding risk regulation regimes*. New York: Oxford Univerty Press, 2004.

IRGC/International Risk Governance Council. *White Paper on Nanotechnology*. Geneva, 2006.

Kolluru, R. Risk assessment and management: a unified approach. In: KOLLURU R.et al. (Org.). *Risk assessment and management handbook*: for environmental, health and safety professionals. Boston: McGraw Hill, 1996. p. 3-41.

Lippmann, M.; Cohen, B. S.; Schlesinger, R. B. *Enviromental health science.* Oxford: 2003.

Lucchese, G. Globalização e regulação sanitária: os rumos da vigilância sanitária no Brasil. 2001. Tese (Doutorado em Saúde Pública) – Escola Nacional de Saúde Pública, Fiocruz, Rio de Janeiro.

National Research Council (United States of America). *Risk assessment in the government: managing the process*. Washington DC: National Academy Press, 1983.

Oberkampf, W. L.et al. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety*, v. 85, p. 11–19, 2004.

Omenn, G. S.; Faustman, E. M. Risk assessment and risk manegement. In: DETELS, Roger et al. *Oxford textbook of public health*. 4 th New York: Oxford University Press, 2005.

Shaw IS, Simões MG. *Controle e modelagem fuzzy*. São Paulo: Editora Edgard Blücher; 1999.

Slovic, P. *The perception of risk*. London: Earthscan, 2000.

Triola, M. F. *Introdução à estatística.* Rio de Janeiro: LTC, 2005.

Vieira Da Silva, L. M. Conceitos, abordagens e estratégias para a avaliação em saúde. In: Hartz, Zulmira M. de A.; Vieira Da Silva, L. M. (Org.). *Avaliação em saúde: dos modelos teóricos à prática na avaliação de programas e sistemas de saúde*. Salvador: Edufba; Rio de Janeiro: Fiocruz, 2005, p. 15 – 39.

Weber L, Klain Pat. Aplicação da lógica *fuzzy* em software e hardware. Canoas: Editora Ulbra; 2003

Zadeh L.A. *Fuzzy* sets. Information and Control 1965; 8:338-353

2

Child Mental Health Measurement: Reflections and Future Directions

Veronika Ottova[1], Anders Hjern[2], Carsten-Hendrik Rasche[1], Ulrike Ravens-Sieberer[1] and the RICHE Project Group[1,3]
[1]*University Medical Center Hamburg-Eppendorf*
[2]*Karolinska Institutet Stockholm University*
[3]*Dublin City University*
[1]*Germany*
[2]*Sweden*
[3]*Ireland*

1. Introduction

Over the course of the past decades, mental health has enjoyed increased interest, particularly in research on subjective health and well-being. In 2008, the EU has launched the European Pact for Mental Health and Well-being in which European Member States declared mental health as an important health issue and recognized it as their responsibility to undertake action. The Pact for Mental Health and Well-being recognizes youth and education as one of the top priority areas for action and sees prevention and reduction of mental disorders (i.e. mental ill-health) as one of the primary objectives (European Commission & WHO, 2008).

According to the World Health Organization's [WHO] definition, health is not "merely the absence of disease or infirmity", but "a state of complete physical, mental and social well-being" (WHO Constitution, 1946). Essential to this definition of health is that it has a positive slant (through the use of the term well-being) and stresses the equal importance of physical, mental and social health. Mental health can further be subdivided into two dimensions: Mental ill-health and positive mental health (Lehtinen et al., 2005). Positive mental health is a resource and is essential to subjective well-being (Lehtinen et al., 2005). Frequently, however, "mental health" is used when actually referring to "positive mental health" and as a consequence is also often (mis)understood as mental health problems or even as mental health diseases/disorders, and not in the positive sense. The persistence of the negative understanding of mental health is largely due to the fact that past and current epidemiological research largely was based on mental health problems and/or illness (Zubrick & Kovess-Masfety, 2005). Many instruments have been developed focusing on mental health problems, thus capturing non-positive outcomes rather than mental health, as such.

"[W]ith its awareness of human capital and education, [modern society] puts a new emphasis on children as the resource of the future, low fertility strengthens children's position as a scarce future resource" (Frønes 2007, p. 7). Upon the background of an

increasing prevalence of chronic disease and mental health problems – in adults and youth populations alike – research into mental health has become increasingly popular over the past years. The term "new morbidity" has been used to describe the changing morbidity pattern (from acute to chronic disease) and the rise of mental health problems (Palfrey et al., 2005). At the present, disabling mental health problems occur worldwide in 20% of children and adolescents (WHO, 2001). This is an alarming number, especially knowing that mental health problems can have a negative effect on the entire society, with consequences, such as loss of productivity and social functioning (Jané-Llopis & Braddick, 2008). The fact that children and adolescents are affected as well is particularly worrisome. We know for instance that the risk for mental health problems in childhood is higher if there is a lack of resources; in the long run this can have effects even later in life (Jané-Llopis & Braddick, 2008). Many adulthood mental health problems have their roots in childhood (WHO, 2005; Jané-Llopis & Braddick, 2008), and therefore monitoring of mental health in children is a promising strategy, particularly in times of profound societal changes (Mortimer & Larson, 2002). Early detection of problem areas is crucial, and therefore, it is essential that monitoring systems are established based on sound indicators.

It is important to stress that despite the above mentioned negative trends, the overall level of mental well-being in Europe is still high (Jané-Llopis & Braddick, 2008). And thus, it is worthwhile not to limit ourselves to only observing patterns of mental health problems, but to look at the positive side as well, in other words: how is the mental health situation in children and adolescents? How can it be measured adequately in this population group to enable identification (screening) of those with good mental health vs. those who are at risk for poor mental health?

The main objective of this chapter is to give the reader a better understanding and appreciation of child mental health measurement, its current state-of-the-art, and additionally, to generally raise attention to this important field of public health. Drawing upon the authors' expertise and involvement in child and adolescent mental health research, the chapter will briefly go into the history of positive mental health and well-being, including important concepts and definitions of mental health, well-being and indicators. The heart of the chapter will be on selected indicators of (positive and ill-) mental health and subjective well-being. Although surely not comprehensive in all regards, this chapter provides a solid background on this research field and the current state-of-the-art of child mental health measurement. A brief discussion with an outlook will close the chapter.

2. Conceptualization of mental health in children and adolescents

2.1 Concepts of mental health and well-being

Coming back to the WHO Definition from the introduction which defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO, 2001, p. 1), the relationship between health and well-being becomes evident. There are two important ideas that emerge from this: first of all, we see that "mental health is an integral part of health, mental health is more than the absence of mental illness, and mental health is intimately connected with physical health and behaviour" (WHO, 2005, p. 2). From this perspective one can also see that "mental health is the foundation for well-being and effective functioning for an individual and for a community" (WHO, 2005, p. 2).

Common terms often used in association with mental health are “emotional and behaviour problems”, “Mental health problems”, “Children’s well-being”, “Psychological health”, “Health-related quality of life”, “Behavioural problems”, just to name a few (Ravens-Sieberer et al., 2008a). However, several of these are synonyms of “ill-mental health” and not terms that in any way describe positive mental health. According to the two continua model, mental health and mental ill-health (mental illness) are related but distinct dimensions (Westerhof & Keyes, 2010). Mental health is a positive phenomenon (Westerhof & Keyes, 2010) and is to be distinguished from ill-mental health.

As a matter of fact, the definition of positive mental health has a long history and goes back to the two traditions of well-being (hedonic well-being and eudaimonic well-being). According to Keyes (2002), good mental health consists of three components which are: emotional well-being (e.g. feelings of happiness and satisfaction with life), psychological well-being (e.g. positive individual functioning in terms of self-realization), and social well-being (e.g. positive societal functioning in terms of being of social value). He extends previous works of Ryff (1989) (six dimensions of psychological well-being) by adding five elements of social well-being which includes “optimal social functioning of inidividuals in terms of their social engagement and societal embeddedness” (Westerhof & Keyes, 2010, p. 111). According to Keyes, positive mental health consists of hedonic well-being and psychological and societal elements of eudaimonic well-being (Keyes, 2005, 2007; Westerhof & Keyes, 2010).

In the following, when using the term well-being, we use the term subjective well-being which is based on self-reports of happiness and life satisfaction (Schwarz & Strack, 1999).

2.2 Historical development: Positive mental health and well-being

Today's understanding of mental health and well-being is the result of scientific research and political activities over the past decades. Building upon the original definition of health (WHO, 1948), the definition of mental health is specified as "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community" (WHO, 2005, p. 2). This contemporary definition contains also the positive view of mental health which is a precondition for well-being. In the eyes of Keyes (2006) the science on mental health and subjective well-being has now arrived at - after half a century starting with the work of Jahoda (1958) on positive mental health - its "third generation" of research which does not merely focus on the absence of illness but "also on the presence of subjective well-being" (Keyes, 2006, p. 1).

The concept of well-being first emerged in Greek philosophical writings. Already ancient civilizations considered health and well-being as one of their highest goods and values in life (Sigerist, 1941). However, the scientific interest in well-being did not begin until the 1950s, when the first indicators for quality of life were defined by social scientists to assess social change and to develop social policy (Land, 1975). The theories emerged during the recreation period after World War II where the "individual's perceptions and viewpoints, and the personal meaning and concerns about life" gained relevance in different scientific fields (Keyes, 2006, p. 2). Especially in philosophy (e.g. Phenomenology, cf. Husserl, 1913), sociology (e.g. Symbolic Interactionism, cf. Blumer 1962), and psychology (e.g. cognitive Psychology, cf. Neisser, 1967), as well as in humanistic theories (cf. Rogers, 1951; Maslow, 1968).

Sponsored by the Joint Commission on Mental Illness and Health in the United States, Marie Jahoda (1958) and Gurin et al. (1960) published two seminal reviews which Keyes categorize as the first research generation of subjective well-being (Keyes, 2006). Jahoda's review 'Current Concepts of Positive Mental Health' (1958) can be seen as pioneer work which shaped our current understanding and theories of positive mental health. Her work was later continued by other scientists like Carol Ryff (1989) who operationalized her theories on well-being (Keyes, 2006). In the first part of this important volume, Jahoda outlines the former understanding of mental health and emphasizes "that the absence of disease may constitute a necessary, but not a sufficient, criterion for mental health" (Jahoda, 1958, p. 15). Further, she investigates throughout a literature research six partly overlapping approaches to categorize positive mental health. They can be summarized as "Attitude of an individual towards his own self", "Self-actualization", "Integration", "Autonomy", "Perception of reality", "Environmental mastery". Yet, there is some criticism especially on the role of cultural influences affecting the understanding of mental health. Scientists, e.g. Murphy (1978), argued that western cultures are predominated by individualism and so other cultures which have a strong collectivistic viewpoint could have a different understanding of mental health. Therefore, cultural values have a strong influence on concepts of mental health (WHO, 2005). The second influential volume is an interview survey on approximately twenty-five hundred Americans conducted by Gurin, Veroff, and Feld (1960) covering the subjective dimension of mental health. Additionally, the volume "featured the hedonic stream of subjective well-being" (Keyes, 2006, p. 3) which together with the eudaimonic stream became more important and dominant in the "second generation" of research (Keyes, 2006). Hedonic well-being can be seen as a part of subjective well-being focusing predominantly on happiness and interest as well as satisfaction with life (Keyes, 2007). More generally contemplated is the existence of positive and the absence of negative affect (Deci & Ryan, 2008) and matches up with our everyday understanding of the word happiness (Waterman, 1993). In contrast, eudaimonia is a feeling of personal expressiveness, self-realization and life satisfaction (Waterman, 1993; Deci & Ryan, 2008). The latter tradition has lost importance in recent well-being research, but contributes important aspects to the concept of well-being (Deci & Ryan, 2008). Three articles stimulated and leveraged the research during this time: Schwartz & Clore (1983) studied how current mood states can affect judgments of happiness and satisfaction with life; Diener (1984) reviewed the first generation of subjective well-being with a focus solely on the hedonic streaming; Ryff (1989) operationalized different aspects of well-being. While the well-being theories became more and more elaborated, various multidimensional scales were developed to measure different aspects of the concept. Early scales for adults included, e.g. the Bradburn Schedule (1969) and the General Well-Being Scale (GWBS) (1969). The first signs of child and adolescent well-being measurement can be found in the "social indicators movement" in the 1960s. Seminal work published by Campbell and Converse (1972) deals with the development of subjective indicators of the quality of life (e.g. aspiration, expectations, and life satisfaction) and Sheldon and Moore's (1968) volume "Indicators of Social Change" conceptualized "objective measures, reviewing available data, and recommending data needs that would enable descriptive reporting on the status of society across domains" (Lippman, 2007, p. 40; Aborn, 1985). In later years, theoretical, normative and methodical changes in science spurred and formed the development of child indicators. Of particular relevance were the United Nation's Convention on the Rights of the Child (CRC) which raised a normative framework for an integral view on children; the "new" sociology of childhood considered it as an independent stage in and of his self and child development theories became more

dynamic processes interacting with the environment (Ben-Arieh, 2008). These theoretical changes gave rise to new methodological perspectives. To better capture children's own living conditions, especially on terms like mental well-being or peer-relations, subjective reports and child-centred indicators became necessary (Ben-Arieh, 2008).

Efforts to synthesize data into national and international "state of the child" reports began during the 1970s. At the international level, UNICEF published in 1979 the State of the World's Children Report (United Nations International Children's Emergency Fund [UNICEF], 1979) – at that time only basic survival indicators were included (e.g. infant and child mortality). In the 1990s, significant developments were made in the reporting. For instance, the Census Bureau published the first comparable report at an international level including domains on family structure, economic status, health and education (Lippman, 2007). Of particular note is the work of an international group of child health experts on a project called "Measuring and Monitoring Child Well-being: Beyond Survival" (Ben-Arieh & Wintersburger, 1997) with the intention to create international indicators which measure quality of life from a child's perspective, including indicators which go beyond the traditionally used survival indicators, such as social connectedness, civic and personal life skills and children's subculture (Lippman, 2007). Until now, a varying set of indicators exist to measure different aspects of child well-being. Crucial factors for the development of mental health indicators were obtained in the child indicators movement: Indicators for negative or risk factor were complemented by indicators for protective factors. Also, the indicators shifted from "well-becoming" (e.g. preparing children to be productive and happy adults) to "well-being" (Ben-Arieh, 2008).

As this short historical overview shows, there have been groundbreaking developments on the understanding and conceptualization of mental well-being. In the next two sections we will take a closer look at mental health measurement in children and adolescents. We will begin by briefly highlighting the importance of indicators in health monitoring while also pointing out the conceptual and methodological challenges.

3. Indicators as tools for health monitoring

The term indicator originates from the Latin word "indicator" which means "one who points out" or "indico" (=to point out). Indicators can cover anything from "indices, signs, and symptoms" to "calculated probabilities and systematic measurements" (Frønes 2007, p. 8), and include time and space. Bauer (1966) has referred to (social) indicators as "statistics, statistical series, and all other forms of evidence […] that enable us to assess where we stand and are going with respect to our values and goals" (p. 1). For policy makers, indicators provide valuable information on relevant public health issues, including their trend and direction of change (improvement or worsening) (Lippman, 2007). But also other groups, such as child advocacy groups, researchers, and media use them for various purposes (Ben-Arieh, 2008).

A good example of a widely-known and politically very influential programme is the OECD Programme for International Student Assessment [PISA]. PISA assesses to what extend knowledge and skills essential for participation in society have been acquired by 15-year-olds at the end of their compulsory education (www.pisa.oecd.org/). The PISA indicators of educational success and marginalization are "perhaps the most well-known example of highly elaborated comparative research indicators related to children" (Frønes, 2007, p. 7). The first PISA report in 2000 had a substantial national and international impact and PISA assessment continues to be an important strategy to benchmark improvements in education at international level.

The European Community Health Indicators Project (ECHI) is a similar effort, but has a different focus. Its aim is to lay the foundation for further development of health indicators targeting all population groups, not just school children. The initial projects on European Community Health Indicators (ECHI and ECHI-2) which were conducted between 1998 and 2005 developed ECHI indicator lists which formed the basis for the follow-up work in the ECHIM project. The ECHIM project is part of the European Health Strategy and builds upon the works of ECHI and ECHI-2. It has three main objectives (Kilpeläinen, Aromaa & the ECHIM project, 2008):

- to further develop health indicators (based on ECHI short list),
- to initiate implementation in the EU countries, and
- to enable the establishment of a Health Monitoring System.

Within ECHI, an indicator was defined as a characteristic of an individual, population or environment which is subject to measurement (directly or indirectly) and can be used to describe one or more aspects of the health of an individual or population (quantity, quality and time). According to ECHI recommendations, indicators must fulfil the criteria of validity, sensitivity, comparability (Kramers, 2003).

Despite advances in indicator development through projects such as ECHI, the development of positive mental health indicators for children and young people is really only beginning (Maher & Waters, 2005). While we seek to gain a better understanding of the magnitude of mental health problems in children, we seem to oversee the importance of measurement tools and indicators to facilitate this process. Monitoring of both positive mental health and mental ill-health (i.e. mental health problems) is essential for human development (Zubrick & Kovess-Masfety, 2005). Unfortunately, mental health research in children and adolescents currently lacks well-established indicators. It is primarily "needs driven", focusing on "illness" rather than "wellness", and in consequence, aimed at physical rather than mental health (Zubrick & Kovess-Masfety, 2005). Furthermore, it is too focussed on distress, and mental health problems, such as delinquency, suicide, depression (Maher & Waters, 2005), rather than positive mental health.

Presently, existing indicators on health are available through organisations, such as the European Union [EU], the Organisation for Economic Co-Operation and Development [OECD], and the World Health Organisation [WHO]. The EU sustainable development indicators provide 120 indicators, the OECD social indicators have 34 indicators on employment, society, general health and social cohesion, and the EU social protection indicators comprise 11 primary social protection indicators (whereby none on mental health). In 2009, the Innocenti Research Centre of the UNICEF has published a working paper on "Positive indicators of child well-being: a conceptual framework, measures and methodological issues" outlining frameworks for further development of positive indicators of well-being of children as well as the challenges involved (Lippman et al., 2009).

4. Child mental health measurement

4.1 Indicators of mental health

As mentioned in the introduction of this chapter, we have limited ourselves to a narrow selection of indicators which we consider suitable for several reasons. First of all, all of the indicators presented here are based on tools/instruments assessing the subjective

perspective of the child itself. Secondly, and this is perhaps even more important, the indicators are based on robust, scientifically valid measurement tools. Apart from having been frequently used in research studies in Europe (as well as internationally), the measures also fulfil the scientific criteria for indicators (as proposed by the ECHI group).

When child well-being is of interest, the preferred method of assessment is via the child's own subjective perspective (Lippman et al., 2009). How children and adolescents reflect and perceive their world and life may differ quite substantially today from adult's reality (Bradshaw et al. 2006). Increasing their participation and asking for their insight and view is an indispensable component of present and future research. "Current attempts to measure children's well-being are problematic because they fail to incorporate an analysis of broader contextual structural and political factors" (Morrow & Mayall, 2010, p. 162). Subjective indicators reflecting the "voice of the child" need to be complemented by objective models on well-being and indicators (Frønes, 2007, p. 11). Furthermore, indicators should be based on measurement tools, which have undergone extensive piloting and ideally have been used previously in surveys. Measurement tools need to be age-, gender-, and culturally-sensitive and should also take the individual's socioeconomic background into account (Erhart et al., 2006).

Many of the indicators which will be presented here originate from the KIDSCREEN survey ["Screening for and Promotion of Health-Related Quality of Life in Children and Adolescents - A European Public Health Perspective"] and the Health Behaviour in School-aged Children [HBSC] Survey.

4.1.1 The KIDSCREEN survey

The European KIDSCREEN project titled "Screening for and promotion of health-related well-being in children and adolescents: a European public health perspective (KIDSCREEN)" took place between 2001 and 2004 in 13 European countries (Austria, the Czech Republic, France, Germany, Greece, Hungary, Ireland, Poland, Spain, Sweden, Switzerland, the Netherlands and the United Kingdom) and had the aim to develop a standardised screening instrument for quality of life in children and adolescents. This instrument should be suitable for representative national and European health surveys and enable cross-cultural comparisons. The project, which also comprised data collection from large population-based samples in each of the participating countries, was part of the Quality of life and Management of Living Resources programme and was funded by the European Commission (EC) within the Fifth Framework Programme (EC Grant Number: QLG-CT-2000- 00751) (Ravens-Sieberer et al., 2001). The data collection targeted children between 8 and 18 years of age using both parents as well as children as information sources. The same kind of data collection tools (questionnaires) and the same assessment tools were used in all participating countries (Ravens-Sieberer et al., 2005). Data on physical health, mental health and socioeconomic status in children and adolescents in Europe was collected and the distribution of mental ill health and poor mental well-being estimated. Thee important instruments came out of the KIDSCREEN project: KIDSCREEN-52, KIDSCREEN-27 and KIDSCREEN-10. Single dimensions of these instruments and the global HRQoL score (KIDSCREEN-10) can be used as suitable indicators for quality of life resp. positive mental health. The KIDSCREEN-10 Mental Health Index assesses the child's perspective on his or her physical, mental and social well-being, identifies children at risk and suggests suitable early interventions. For this

reason it is particularly useful for identifying children with positive mental health. Section 4.4 of this chapter will present empirical results from the KIDSCREEN survey detailing the distribution of children with positive mental health in thirteen European countries.

Further information on the KIDSCREEN instruments is available at http://www.kidscreen.org.

4.1.2 The Health Behaviour in School-aged Children (HBSC) survey

The HBSC Study is a WHO-collaborative study dedicated to the study of adolescent health. The overall aim is to gain a better understanding of health behaviours, health, and well-being in children and adolescents at the age of 11, 13 and 15 years. HBSC is a cross-national study covering over 40 countries in Europe, North America and Israel. The design of the survey is cross-sectional and data collection is carried out every four years. The basis for each survey is a standardized research protocol which is renewed for each survey round. The survey is based on a questionnaire which consists of mandatory items (required from each country), and optional items which focus on topics of national interest. Mandatory items are part of the international file and enable cross-country comparisons. Data is collected in schools and the primary sampling unit is school class (or entire school in case this is not possible). Data is collected within a class period via questionnaire.

Further information on the HBSC Survey is available at: http://www.hbsc.org .

Collection of data on positive mental health is in line with the health definition of the WHO (Ravens-Sieberer et al., 2008a). Assessment of mental health is possible in one of two ways: positive mental health and negative (ill) mental health, and the HBSC and KIDSCREEN Surveys provide suitable instruments for both.

4.2 Positive mental health indicators

4.2.1 Quality of life and positive mental health indicator

One of the outputs of the European KIDSCREEN survey was the development of a screening tool for mental health. The KIDSCREEN-10 instrument is an index which assesses the child's perspective on his or her physical, mental and social well-being, thus enabling the identification of children at risk. As previously stated, the KIDSCREEN-10 Mental Health Index is a non-clinical measure of quality of life and positive mental health status and enables the assessment of school-aged children's general well-being. The index is especially sensitive for affective, cognitive, and psychovegetative, as well as psychosocial aspects of mental health.

The short instrument consists of ten items covering six aspects of quality of life (physical well-being, moods & emotions, autonomy, parent relation & home life, peers & social support, school environment). The short instrument consists of the following ten items and only takes a few minutes to complete:

- "Have you felt fit and well?"
- "Have you felt full of energy?"
- "Have you felt sad?"
- "Have you felt lonely?"

- "Have you had enough time for yourself?"
- "Have you been able to do the things that you want to do in your free time?"
- "Have your parent(s) treated you fairly?"
- "Have you had fun with your friends?"
- "Have you got on well at school?"
- "Have you been able to pay attention?"

Children at risk of poor quality of life are identified by coding of responses so that higher values indicate better quality of life. The KIDSCREEN-10 Mental Health Index was developed by means of a Rasch analysis which ensured that only those items which represented a global, unidimensional latent trait were included. The values on the individual items are summed up, Rasch person parameters (PP) are assigned to each possible sum score, and then the PP are transformed into values with a mean of 50 and standard deviation of approximately 10 (Ravens-Sieberer et al., 2006). A better differentiation between the children is made possible by the distribution of the Rasch scores that resemble the expected theoretical normal distribution. The index provides a good discriminatory power and shows only few ceiling or floor effects.

Validation work on this instrument indicates that it is a valid and well-tested stable child centred self-report measure (indicator) for child and adolescent general quality of life and mental well-being status. It has good psychometric properties, with high reliability and Rasch-scale properties. The index provides a good discriminatory power and shows only few ceiling or floor effects. The strong internal consistency reliability (Cronbach's Alpha = .82) and test-retest reliability (r = .73) allow precise and stable measurements (Ravens-Sieberer et al., 2006; Ravens-Sieberer et al., 2010).

The cut-off at T-value below 38 (which represents the lowest 10%) indicates lower quality of life resp. higher risk for poor mental health (Ravens-Sieberer et al., 2006). Comparisons with the European Community Health Indicators (ECHI) show that the Kidscreen-10 Index for children and adolescents corresponds well with the "General Quality of Life Indicator". The ECHI group proposes to use the Euroqol score from the Euroqol 5D instrument (Eurociss project) or alternatively the WHOQOL of the WHO (Kramers & the ECHI team, 2005) for adults.

Since its development, the instrument has been employed in several EU funded European research projects (KIDSCREEN, DISABKIDS, MHADIE, SPARCLE), in the Flash Eurobarometer and in the PROMIS roadmap initiative of the US NIH (National Institutes of Health) to develop a patient reported outcome measurement information system. The instrument is also used in the Health Behaviour in School-aged Children (HBSC) study as an indicator of positive mental health and has been translated into a variety of languages.

4.2.2 Psychological well-being indicator

Another positive mental health indicator is psychological well-being which refers to a child's or adolescents' positive emotions and perceptions, his/her satisfaction with life, covering various areas of his/her inner feelings and thus provides insight into an individual's mental health state. The psychological well-being dimension is one of ten dimensions of KIDSCREEN-52 and one of the five dimensions of KIDSCREEN-27 as shown in the figure below. In the latter, it also encompasses the Moods and Emotions and the Self-Perception scale of KIDSCREEN-52.

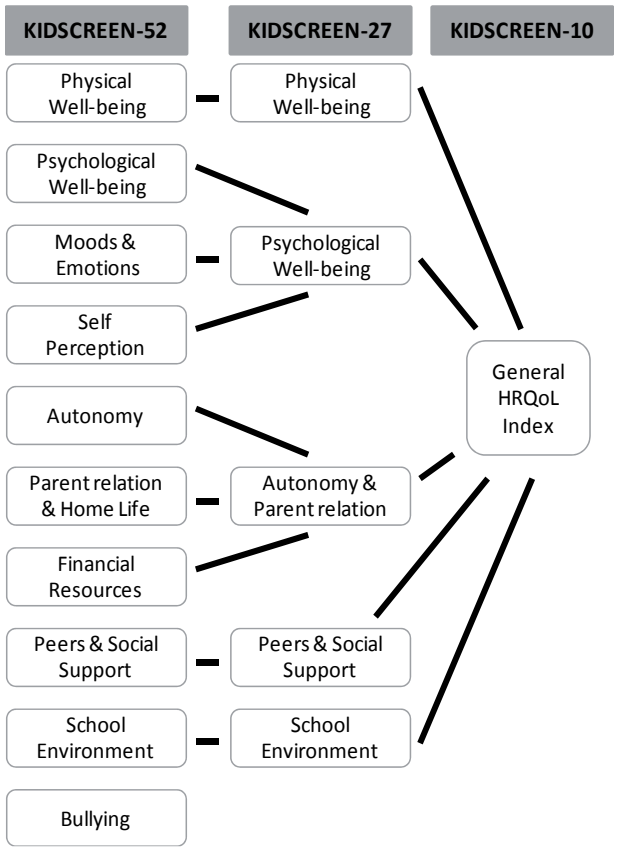


Fig. 1. Dimensions of the KIDSCREEN-52, -27, -10 (http://www.kidscreen.org/)

The KIDSCREEN-52 is part of a family of health-related quality of life instruments which were developed in several stages, beginning with literature searches, expert consultations (Delphi method) and focus groups with children and adolescents (Herdman et al., 2002; Ravens-Sieberer et al., 2006, 2008b). Using this approach, relevant health-related quality of life (HRQoL) dimensions and items could be identified (Ravens-Sieberer et al., 2006, 2008b). Reduction of items gathered in focus groups were done following EUROHIS guidelines (Nosikov & Gudex, 2003). Following this, a procedure of forward-backward-forward translation and harmonization was applied, followed by a pilot study and an item reduction analysis, which finally yielded a questionnaire comprising 52 items (Ravens-Sieberer et al., 2006, 2008b).

The KIDSCREEN-52 Psychological Well-being dimension assesses the psychological well-being of the child, which covers positive emotions and life satisfaction, including the child's or adolescents' positive perceptions and emotions, and positive feelings, such as happiness, joy and cheerfulness (Ravens-Sieberer et al., 2006). A low score on this dimension implies no pleasure in life/high dissatisfaction with life while a high score indicates happiness, positive view of life, life satisfaction and cheerfulness. The cut-off is at T-value of 36.91 and identifies the lowest 10% (Ravens-Sieberer et al., 2006). This dimension of the KIDSCREEN-52 (for children and adolescents) is comparable to the psychological well-being dimension for adults which is used as an indicator for general mental health in the ECHI report, and is

defined as the percent of the population below the cut-point of the energy-vitality scale from the SF-36 questionnaire (see ECHI long list, 2005).

4.2.3 Self-rated health (subjective health indicator)

The building blocks for good health are laid early in life, and therefore evaluation of health does not begin in adulthood but much earlier. Health is an important resource and poor health early in life can have long-term negative effects which may continue throughout adulthood (WHO, 2006). Being in good health – physically, emotionally and socially – helps young people deal productively with challenges in their development (Burt, 2002). In recent years, self-assessments of health have come more into use as they are based on an individual's perception and evaluation of his or her health. Focusing on the subjective perspective, self-rated health is usually founded on age-peer comparisons either consciously or unconsciously (Bjorner et al., 1996). It can be distinguished from more specific health constructs in that it captures an overall conception of health, rather than a summation across specific domains of health. Empirical studies have shown that self-reported health is an independent predictor of mortality (Idler & Benyamini, 1997). Benjamins et al. (2004) could also identify a relationship between self-reported health and cause-specific mortality, and moreover, also found gender effects for some causes of mortality. A gender effect in self-rated health was confirmed in a sample of children, whereby girls reported poorer health than boys (Cavallo et al., 2006). Another study on psychosocial, demographic, and health-related correlates of self-rated health showed that daily smoking, alcohol intoxication on at least one occasion, infrequent physical activity, and difficulty making friends were predictors of poor self-rated health (Kelleher et al., 2007). Seemingly, multiple independent correlates of adolescent self-rated health exist (Breidablik et al., 2009), whereby poor health increases by age and throughout adolescence (Wade & Vingilis, 1999).

The single item question on health is a suitable indicator of subjective health. Individuals are asked to indicate how they perceive their general health on a Likert scale. The answer categories are either four or five scaled. The Health Behaviour in School-aged Children (HBSC) study uses the four point Likert scale with answer categories: "excellent", "good", "fair", "poor". Those with either "fair" or "poor" health respectively "excellent" and "good" health are then combined into subgroups of individuals with "poorer health" resp. "better health" (Currie et al., 2004). Other studies, such as e.g. the European KIDSCREEN survey (Ravens-Sieberer et al., 2006), use the five-scaled answer categories: "excellent", "very good", "good", "fair", "poor". The main difference is the extra answer category "very good" which enables more differentiation on the positive end of the scale.

According to ECHI recommendations, preference should be given to the five answer categories (Kramers & the ECHI team, 2005). In the ECHI report, the WHO recommended instrument is proposed which is based on a five response category item: "How is your health in general?" (Answer options are: very good/good/fair/bad/very bad). The ECHI report further proposes to set the cut-off for perceived health at the % (very) good/less than good/less than fair. As noted in the ECHI report, very little focus was placed on the specific situation of children (ECHI long list 2005, p. 50).

Although no specific validation has been done on the self-rated health item in HBSC, several studies support its validity. It has shown multiple independent health-related

correlates, including medical diagnosis, and health complaints. The self-rated health item shows a certain degree of stability across time, suggesting that these self-reports are not simply a fluctuating subjective impression. Cavallo et al. (2006) analyzed the item in terms of its feasibility and psychometric robustness using the HBSC 2001/2002 data from all countries involved. The results confirm the trend of an increasing perception of poor health with increasing age in the pre-adolescence phase and a higher risk for perceived poorer health in girls, (Cavallo et al., 2006). HBSC showed this was a consistent finding across a large number of countries in Europe and North America (see also Currie et al., 2008).

4.2.4 Life satisfaction indicator

Well-being is a multi-faceted concept (Diener, 1984; Wilkinson & Walford, 1998) and comprises the individual's own evaluation of life, i.e. life satisfaction. It was not until the early 1990s that determinants of life satisfaction were studied (Suldo et al., 2006). Unlike other concepts, life satisfaction is relatively stable over time (Pavot & Diener, 1993). It is associated with depression, anxiety, suicide, work disability, fatal accidents and all cause mortality in adults (Fiscella & Franks, 1997; Helliwell, 2007; Koivumaa-Honkanen et al., 2001; Koivumaa-Honkanen et al., 2002; Koivumaa-Honkanen et al., 2004a; Koivumaa-Honkanen et al., 2004b). During adolescence, life satisfaction is influenced by life experiences and relationships, especially within the family context (Edwards & Lopez, 2006; Gohm et al., 1998; Rask et al., 2003) and school (Samdal et al., 1998). Psychosocial resources and school satisfaction, especially perceptions of feeling treated fairly, feeling safe and perceiving teachers as supportive (Samdal et al., 1998), are linked with high life satisfaction. School-related resources and their impact on overall life satisfaction are a central issue as the acquirement of academic competence constitutes one of the developmental goals in adolescence (Hurrelmann & Lösel, 1990). Moreover, school creates a social environment for young people which can provide them with additional resources. At the certain time, some social factors, such as bullying, can pose a risk as they may be associated with low life satisfaction and low subjective health (Gobina et al., 2008).

The Cantril Ladder is a measure of life satisfaction which has been widely used in the Health Behaviour in School-aged Children (HBSC) study. The measure is also a suitable indicator for life satisfaction in children and adolescents (Cantril, 1965). The measure consists of a Visual Analogue Scale with 11 positions (0 through 10) where children can mark the position on the scale demonstrating how satisfied they are with their life: "Here is a picture of a ladder. The top of the ladder "10" is the best possible life for you and the bottom "0" is the worst possible life for you. In general, where on the ladder do you feel you stand at the moment? Tick the box next to the number that best describes where you stand." The Health Behaviour in School-aged Children (HBSC) study uses the cut-off at "6 or above" to identify children and adolescents with a positive level of life satisfaction (normal to high life satisfaction) (Currie et al., 2004).

The Cantril Ladder has not been subject to structured validation studies at the international level, but observed relationships with quality of life and with self-rated health are in the expected range, and support claims about its validity. Analyses using data from the HBSC study show that the item is associated with the general health item and the Symptom Checklist (HBSC-SCL) (Cavallo et al., 2006).

4.3 Ill-mental health indicators

4.3.1 Psychological distress indicator

In the past, assessment of mental health was for the most part aimed at assessing mental ill health, with the focus being placed on mental health disorders and -problems. This has the disadvantage that the information gathered only enables separation between individuals with (signs of) mental disorders and healthy individuals (without any signs of mental health problems). No information is available on individuals "in-between", in other words about the position of the individual on a mental health continuum (Ravens-Sieberer et al., 2008a). Moreover, earlier instruments for measuring mental health problems in children were based and validated on experiences with child psychiatric patients and were often developed as screening instruments for patients in care. KIDSCREEN instruments overcome this drawback as they have been developed to measure mental health in the general population and have been validated in large population studies.

The KIDSCREEN-52 "Moods & Emotions" Dimension provides an important indicator of psychological distress which can be used to identify children with depressiveness, as well as those feeling lonely, sad, and unhappy (Ravens-Sieberer et al., 2006). This dimension of the KIDSCREEN-52 examines experiences of depressive moods and emotions, including stressful feelings, and how distressing these are to the individual. A low score indicates that the child or adolescent feels depressed, is unhappy and/or in bad mood. A high score in contrast, implies feeling good and being in a good mood. The cut-off identifying the lowest 10% is at a T-value of 37.76 (Ravens-Sieberer et al., 2006).

The "moods & emotions" dimension of KIDSCREEN-52 for children and adolescents corresponds to the indicator of psychological distress for adults in the general mental health section as published in the ECHI indicator list. In the ECHI report, psychological distress is defined as the percent of the population below the cut-point of MHI-5 score from the SF-36 questionnaire (see ECHI report, long list: http://www.echim.org/docs/echi_longlist.pdf).

The KIDSCREEN instruments are robust and psychometrically sound instruments suitable for the assessment of the health-related quality of life and mental health in children and adolescents between 8 and 18 years of age. The internal consistency reliability (Cronbach's Alpha) for the individual dimensions show for the "Moods & Emotions" dimension a value of 0.86 and for the "Psychological Well-being" dimension a Cronbach's Alpha of 0.89 (Ravens-Sieberer et al., 2008b; Ravens-Sieberer et al., 2006), both of which can be considered sufficiently high.

Both, the "Moods & Emotions", as well as the "Psychological Well-being" dimension of KIDSCREEN-52, correspond well with the published indicators of general mental health in the ECHI report (Kramers & the ECHI team, 2005), and are thus suitable indicators.

4.3.2 Subjective health complaints index

The presence of subjective health complaints and the frequency of their occurrence can serve as a good approximation for the individual's physical well-being. Health complaints tend to cluster together (Alfven, 1993; Mikkelsson et al., 1997; Starfield et al., 1984; WHO, 2006) and in this way cause immense burden – not only on the individual, but also on the health care system.

Within the international HBSC Study, the Symptom Checklist (HBSC-SCL) was developed to assess the various health complaints that might occur in children and adolescents. The

HBSC-SCL has proven to be a suitable and effective screening tool for the assessment of physical well-being. The Checklist includes symptoms, such as headache, abdominal pain, backache, feeling low, irritability or bad mood, feeling nervous, sleeping difficulties and dizziness (Haugland et al., 2001). The advantage of the HBSC-SCL is that it is not limited to somatic symptoms, but also contains a number of psychological symptoms and hence constitutes an instrument suitable for detecting psychosomatic complaints.

The HBSC-SCL assesses the occurrence of health complaints in children and adolescents and is a useful indicator for identifying individuals at risk for impaired health. The HBSC-SCL asks about the occurrence of the following symptoms in the last 6 months: Headache, Stomach ache, Back ache, Feeling low, Irritability or bad temper, Feeling nervous, Difficulties in getting to sleep, Feeling dizzy. Ravens-Sieberer et al. (2008c) developed an international scoring system for the HBSC-SCL which enables a cross-cultural and interval-scaled assessment of subjective health complaints and which can further be used to identify individuals at a greater risk of health complaints. This uni-dimensional scoring algorithm is based on seven of the eight items. A score below 41 indicates a "higher risk" for health complaints (Ravens-Sieberer et al., 2008c).

A number of validation studies have been made on the HBSC-SCL (Haugland & Wold, 2001; Haugland et al., 2001). Qualitative semi-structured interviews with early adolescents revealed that adolescents perceive the symptoms to be aversive physical and psychological states that interfere with daily functional ability and well-being. Consistent accounts as to how the different symptoms were defined were given, suggesting that adolescents have a common frame of reference when they rate their frequency of symptoms (Haugland & Wold, 2001). Differences emerged in their lay perspectives on the causes of such symptoms. While some explanations were consistent with a stress-model of health complaints, others were associated with developmental processes, such as growing pain, or ergonomic factors, such as low quality of air in classrooms etc.

4.4 Application of the mental health indicator (KIDSCREEN-10)

As previously mentioned, the mental health index (KIDSCREEN-10) is a non-clinical measure of mental health status. It does not permit identification of groups with defined burden of mental health problems, but allows measurement along a continuum (Ravens-Sieberer et al. 2008a).

As stated previously, KIDSCREEN-10 is an indicator of quality of life and (positive) mental health and in the following, we will apply it on a European sample of adolescents from 13 countries. The overall mean score is 48 with a standard deviation of 10. The results in Figure 2 show that some countries fall above and some below the European mean. Countries towards the left side of the figure tend to show better positive mental health compared to the countries at the right end of the figure which fall below the European average. Additional results show that variation in mental health scores was generally lower in countries with lower positive mental health scores (results not shown).

To gain a better understanding of the distribution of mental health, we will now look at selected sociodemographic characteristics, such as gender and socioeconomic status (approximated by family affluence [FAS]). Table 1 below shows the distribution of positive mental health across the 13 countries by gender. Comparisons across gender groups show that boys report better mental health across all countries than girls. This difference is significant in all but one country, and the effect sizes are generally small.

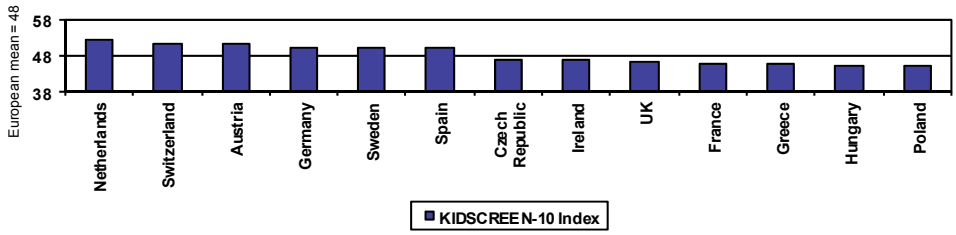


Fig. 2. Positive mental health index (KIDSCREEN-10) across 13 European countries[1,2]

Country	Girls m(SD)	Boys m(SD)	Effect (d)
Austria (n=878)	49.6 (8.9)	52.6 (9.4)	0.3***
Czech Republic (n=1016)	45.0 (7.1)	47.3 (7.8)	0.3***
France (n=622)	45.0 (8.4)	46.1 (8.0)	n.s.
Germany (n=1079)	49.3 (8.4)	51.0 (8.4)	0.2***
Greece (n=1146)	44.2 (7.6)	47.2 (8.0)	0.4***
Hungary (n=1839)	43.6 (7.6)	46.2 (8.9)	0.3***
Ireland (n=894)	45.5 (7.9)	48.1 (7.6)	0.3***
The Netherlands (n=1168)	50.2 (8.2)	53.6 (10.0)	0.4***
Poland (n=1120)	43.9 (7.9)	45.3 (7.3)	0.2**
Spain (n=522)	48.4 (9.6)	50.9 (8.7)	0.3**
Sweden (n=3097)	49.2 (10.0)	52.4 (10.0)	0.3***
Switzerland (n=1078)	49.6 (8.0)	52.6 (8.5)	0.4***
United Kingdom (n=883)	45.5 (8.3)	47.8 (8.5)	0.3***

** p<.01
*** p<.001

Table 1. Positive mental health (KIDSCREEN-10) in different countries according to gender[3,4]

Next, the analysis of positive mental health by family affluence shows that higher family affluence, i.e. growing up a in a better-situated family, is generally associated with a higher level of positive mental health (i.e. above the European average). This is depicted in Figure 3 by the increasing line (with one or two exceptions) and also in the distribution of % of children in low, middle and high FAS group per country (results not shown). Countries with higher mental health score means are also those with the least number of adolescents in the low FAS group.

[1] Mean scores of the KIDSCREEN-10 are depicted
[2] This figure was previously published in Ravens-Sieberer et al. (2008a).
[3] Effect size calculation was based on dividing the mean difference by the overall standard deviation (according to Cohen 1988).
[4] This figure was previously published in Ravens-Sieberer et al. (2008a).

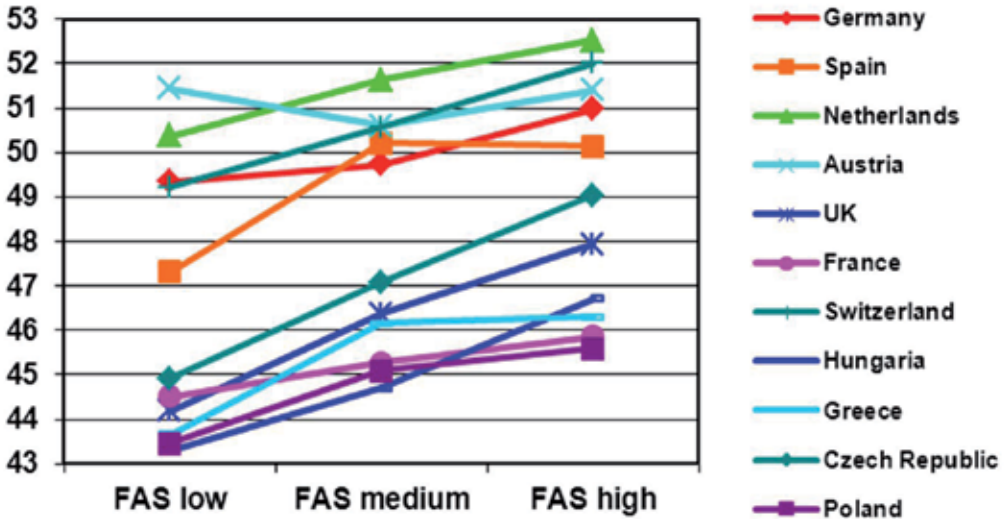


Fig. 3. Adolescent positive mental health (KIDSCREEN-10) and FAS in the participating countries (without Ireland and Sweden)[5]

The results that were presented here serve the purpose to exemplify the application of a robust measure of mental health in a European sample of children and adolescents. Results show the associations between the outcome (mental health) and various sociodemographic factors (age, gender, FAS) and in this way provide the basis for more comprehensive analyses of mental health status in children and adolescents in Europe.

5. Closing comments

The objective of this chapter was to give interested readers an insight into the state-of-the-art in child mental health measurement. Our aim was to show that progress has indeed been made in this vast field, and although we still do not have all the tools and information for a complete assessment of mental health in children and adolescents, we have been able to identify useful measures and important surveys at the European level which enable a good approximation. The complexity of the field has made it necessary for us to concentrate on a few indicators which in our view are good representatives of the respective constructs. The indicators and the results we have described in this chapter come from the HBSC and KIDSCREEN Studies and also reflect our insights from within the RICHE project.

The original idea for a publication on mental health measurement came up during the course of working in the RICHE project. RICHE stands for "Research into Child Health in Europe" and is an international project focusing on child health research in Europe. The project is funded within the EU 7th Framework Programme. RICHE embraces the full multi-disciplinary diversity of European research and addresses its fragmentation by making the parts visible. This is done in part via a platform which provides the opportunity for open exchange (http://www.childhealthresearch.eu/). The aims of the project are: to provide an inventory of current research; to identify research into child health measurement, statistics,

5 The Figure was previously published in Ravens-Sieberer et al. (2008a).

and indicators; to identify gaps in child health research as perceived by stakeholders; and lastly, to develop roadmaps for the future of child health research in Europe based on all these findings. One of the objectives of the RICHE project is to produce an inventory of measurement and indicators to facilitate the implementation of existing methods and at the same time to initiate new developments through exchange and networking. Based on the notion that development and implementation of sound indicators is essential for developing child health with the European Union (Rigby et al., 2003), it is important that high quality indicators are available for analyses and political decision making. Health measurement is the core for developing prevention strategies in a life course perspective.

The review of research into child mental health measurement has revealed important advances, such as the development of quality of life and positive mental health indicators in the KIDSCREEN project, while on the other hand, also pointed out several shortcomings. The main shortcomings relate to the age-specification and the cultural adaptation of measurement tools. With regard to the age issue, our evaluation of measures of child mental health revealed that measurement tools generally target older children and adolescents, i.e. eight years and older. Many of the measurement tools come from the HBSC survey and all of these have originally been designed for 11-, 13-, and 15-year old children. Any application of these instruments on younger children would require further validation which to date has not been done. Although indicators based on KIDSCREEN measures are suitable for slightly younger children (beginning with age seven), they are not available for the very young children (0-6 year olds). For this young group, there is a clear gap on measurement tools, especially those enabling a valid and cross-cultural assessment of quality of life and well-being for the age group 0-3 years and for the age group 4-6 years. Generally, very young and young children are underrepresented in international data sources, and "a portrait of positive well-being among young children is not available, and in many cases, measures are lacking that are appropriate for their age" (Lippman et al., 2009, p. 24). This implies that many indicators are adolescent-focused and hence may point attention to matters relevant for adolescents which may be quite different from those that are essential for children (Bradshaw et al., 2006).

Another shortcoming of current research on indicators for child well-being lies in the cultural adaptation of the measurement tools. As mentioned above, all of the indicators presented here are based on measurement tools which have been developed within the European and North American context. In order to compare child well-being and quality of life in different cultural contexts outside of Europe (e.g. in Africa, Asia), cultural adaptations would need to be done with the instruments. Currently, this is a research challenge in this field and needs to be addressed in the near future.

6. Outlook

To end on a positive note, it is important to acknowledge that there are already a number of programmes on mental health and well-being in children and adolescents underway. A good example of a promising strategy is Scotland's "National Programme for Improving Mental Health and Wellbeing" which was launched in April 2008 with the purpose to identify a core set of indicators on mental health to support the national action plan on mental health ("mental health profile for Scotland"; Parkinson, 2009). Building upon the experience from the establishment of a mental health indicator set for adults (Parkinson,

2007), the same is now being done for the age group 18 years and younger. The goal is to "support and promote consistent and sustainable national monitoring of the state of mental health and associated contextual factors for children and young people in Scotland" (Parkinson, 2009, n.p.). Using different methods, such as a comprehensive review of literature on children's own views of mental health (Shucksmith et al., 2009), and wider consultations with researchers, policy makers and practitioners, as well as advisory groups will be used to inform the development of a framework on mental health and well-being. Direct consultations with children and young people on mental health indicators and the proposed framework will complement this information (Parkinson, 2009).

NHS Scotland is a good example of how research and policy making can work together to more forward on an important public health issue with high relevance not just in Scotland, but also at the European level. It highlights the importance of epidemiological data which delivers information relevant for development of policy making (Remschmidt & Belfer, 2005). HBSC and KIDSCREEN are good examples of this. Through regular, standardized data collection (such as through monitoring), health indicators can further help in the problem identification process as well as in its prioritization (Korkeila et al., 2006). By "screening" for certain (risk) groups or health problems, they are valuable tools for preventive action which requires early detection of hidden or manifest mental health problems (Erhart et al., 2009). In this sense, indicators are an important "bridge between health policy and scientific information" (Korkeila et al., 2006, p. 13).

7. Acknowledgement

The RICHE Project (Title: A platform and inventory for child health research in Europe) is financed by a grant from the European Commission within the EC Seventh Framework Programme (HEALTH-2009-3.3-5 European child health research platform).

Participants of the RICHE project are:

Prof Anthony Staines, Dublin City University (Coordinator)
Dr Giorgio Tamburlini, Burlo Garofolo
Ms Csilla Kaposvari, Egészség Monitor
Prof Hein Raat, Erasmus Universitair Medisch Centrum Rotterdam
Dr Else-Karin Groholt, Nasjonalt Folkehelseinstitutt
Prof Margarida Gaspar de Matos, Faculdade de Motricidade Humana
Dr Anne McCarthy, The Health Research Board
Dr Mitch Blair, Imperial College of Science, Technology and Medicine
Dr Matilde Leonardi, Fondazione IRCCS Istituto Neurologico Carlo Besta
Dr Polonca Truden-Dobrin, Institute of Public Health of the Republic of Slovenia
Dr Reli Mechtler, Johannes Kepler University Linz, Institute of Health System Research
Prof Michael Rigby, Nordic School for Public Health
Prof Anders Hjern, Nordic School for Public Health
Dr Polanska Kinga, Nofer Institute of Occupational Medicine
Mr Con Hennessy, Open Applications Consulting Limited
Prof Geir Gunnlaugsson, Reykjavik University
Prof Mika Gissler, National Institute for Health and Welfare
Prof Toomas Veidebaum, National Institute for Health Development

Dr Ales Bourek, Masarykova Univerzita
Prof Candace Currie, The University of Edinburgh, National Institute for Health and Clinical Excellence
Prof Ulrike Ravens-Sieberer, Universitätsklinikum Hamburg-Eppendorf
Prof Allan Colver, University of Newcastle upon Tyne
Prof Livia Popescu, Universitatea Babeş-Bolyai
Prof Angela Brand, Universiteit Maastricht, European Centre for Public Health Genomics

8. References

Aborn, M. (1985). Statistical legacies of the social indicators movement. *Paper presented at the annual meeting of the American Statistical Association*, Las Vegas, Nevada

Alfven, G. (1993). The covariation of common psychosomatic symptoms among children from socio-economically differing residential areas: an epidemiological study. *Acta Paediatr*, 82, pp. 484-487

Bauer, R. A. (Ed.) (1966). *Social Indicators*, MIT Press, Cambridge

Ben-Arieh, A. & Wintersberger, H. (Eds.) (1997). Monitoring and Measuring the State of Children – Beyond Survival. *Eurosocial Report No. 62*, European Centre for Social Welfare Policy and Research, Vienna

Ben-Arieh, A. (2008). The Child Indicators Movement: Past, Present, and Future. *Child Indicators Research*, 1, pp. 3-16, 1874-8988

Benjamins, M. R., Hummer, R. A., Eberstein, I. W. & Nam, C. B. (2004). Self-report health and adult mortality risk: an analysis of cause-specific mortality. *Social Science & Medicine*, 59, pp. 1297-1306

Bjorner, J. P, Kristensen, T. O., Orth-Gomer, K., Tibblin, G., Sullivan, M. & Westerholm, P. (1996). *Self-rated health. A useful concept in research, prevention and clinical medicine,* Forskningsradsnämnden, Ord & Form AB, Uppsala

Breidablik, H. J., Meland, E. & Lydersen, S. (2009). Self-rated health during adolescence: stability and predictors of change (Young-Hunt study, Norway). *European Journal of Public Health*, 91, 1, pp. 73-78

Blumer, H. (1962). Society as Symbolic Interaction, In: *Human Behavior and Social Process: An Interactionist Approach*, Arnold M. Rose, pp. 179-192, Houghton-Mifflin, Boston

Bradburn, N. M. (1969). *The structure of psychological well-being*, Aldine, Chicago

Bradshaw, J., Hoelscher, P., Richardson, D. & UNICEF (2006). *Comparing Child Well-Being in OECD Countries: Concepts and methods. Innocenti Working Paper No. 2006-03*, UNICEF Innocenti Research Centre, 1014-7837, Florence.

Burt, M.R. (2002). Reasons to invest in adolescents. *Journal of Adolescent Health*, 31, 2, pp. 136-152

Cameron, E., Mathers, J. & Parry, J (2006). Health and well-being: questioning the use of health concepts in public health policy and practice. *Critical Public Health*, 16, 4, pp. 347-354, 1469-3682

Campbell, A. & Converse, P.E. (Eds.) (1972). *The Human Meaning of Social Change*, Russell Sage Foundation, New York

Cantril, H. (1965). *The pattern of human concern*, Rutgers University Press, New York

Cavallo, F., Zambon, A., Borraccino, A., Ravens-Sieberer, U., Torsheim, T. & Lemma, P. (2006). Girls growing through adolescence have a higher risk of poor health. *Quality of Life Research*, 15, 10, pp. 1577-1585, 1573-2649

Cohen, J. (1988). *Statistical power analysis for the behavioral science*, Lawrence Erlbaum, New York

Currie, C., Roberts, C., Morgan, A., Smith, R., Settertobulte, W., Samdal, O. & Rasmussen, V. B. (Eds.) (2004). *Young People's Health in Context: international report from the HBSC 2001/02 survey, (Health Policy for Children and Adolescents, No.4)*, WHO Regional Office for Europe, 92 890 1372 9, Copenhagen

Currie, C., Nic Gabhainn, S. Godeau, E., Roberts, C. Smith, R. Currie, D., Pickett, W. Richter, M., Morgan, A. & Barnekow, V. (Eds.) (2008). *Inequalities in young people's health: HBSC international report from the 2005/2006 Survey*, WHO Regional Office for Europe, Copenhagen

Deci, E. L. & Ryan, R. M. (2008). Hedonia, Eudaimonia, and Well-Being: An Introduction. *Journal of Happiness Studies*, 9, pp. 1-11, 1573-7780

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95, pp. 542–575, 0033-2909

Dupuy, H. J. (1977). The General Well-being Schedule, In: *Measuring health: a guide to rating scales and questionnaire (2nd ed)*, McDowell & Newell, pp. 206-213, Oxford University Press, Oxford

ECHI long list (2005). Annex 5, In: *The ECHI comprehensive indicators list (Long List)*, 31.10.2011, Available from: <http://www.healthindicators.eu/healthindicators/object_binary/o2701_ECHI_longlist.pdf>

Edwards, L. M. & Lopez, S. J. (2006). Perceived family support, acculturation, and life satisfaction in Mexican American youth: A mixed-methods exploration. *Journal of Counselling Psychology*, 53, pp. 279-287, 1939-2168

Ehrlich, D. (1961). Americans View Their Mental Health (Review). *Archives of General Psychiatry*, 5, 6, pp. 616-a-618, 0003-990X

Erhart, M., Wille, N., Ravens-Sieberer, U. (2006). Die Messung der subjektiven Gesundheit: Stand der Forschung und Herausforderungen, In: *Gesundheitliche Ungleichheit. Grundlagen, Probleme, Perspektiven*, Richter, Hurrelmann, pp. 321-338, VS Verlag für Sozialwissenschaften, 3531160842, Wiesbaden

Erhart, M. , Ottova, V., Gaspar, T., Jericek, H., Schnohr, C., Alikasifoglu, M., Morgan, A., Ravens-Sieberer, U. & HBSC Positive Health Focus Group (2009). Measuring mental health and well-being of school-children in 15 European countries using the KIDSCREEN-10 Index. *International Journal of Public Health*, 54, 2, pp. 160-166

European Commission & WHO (2008). European Pact for Mental Health and Well-being, *Proceeding of EU High-level Conference: Together for Mental Health and Well-Being*, Brussels, June 2008

Fiscella, K. & Franks, P. (1997). Does psychological distress contribute to racial and socioeconomic disparities in mortality? *Social Science & Medicine*, 45, pp. 1805–1809, 0277-9536

Frønes, I. (2007). Theorizing indicators. On Indicators, Signs and Trends. *Social Indicators Research*, 83, pp. 5-23, pp. 1573-0921, 1573-0921

Gobina, I., Zaborskis, A., Pudule, I., Kalnins, I. & Villerusa, A. (2008). Bullying and subjective health among adolescents at schools in Latvia and Lithuania. *International Journal of Public Health*, 53, 5, pp. 272-276, 1661-8556

Gohm, C., Oishi, S., Darlington, J. & Diener, E. (1998). Culture, parental conflict, parental marital status, and the subjective well-being of young adults. *Journal of Marriage and the Family*, 60, pp. 319–334, 1741-3737

Gurin, G., Veroff, J. & Feld, S. (1960). *Americans View Their Mental Health*, Basic Books, New York

Helliwell, J. F. (2007). Well-being and social capital: Does suicide pose a puzzle? *Social Indicators Research*, 81, pp. 455-496, 1573-0921

Herdman, M., Rajmil, L., Ravens-Sieberer, U., Bullinger, M., Power, M., Alonso, J., the European KIDSREEN Group & DISABKIDS Group (2002). Expert consensus in the development of a European health-related quality of life measure for children and adolescents: a Delphi study. *Acta Paediatrica*, 91, 12, pp. 1385–90, 1651-2227

Hurrelmann, K. & Lösel F. (1990). Basic issues and problem of health in adolescence, In: *Health hazards in adolescence*, Hurrelmann & Lösel, pp. 1-21, Walter de Gruyter, Berlin

Husserl, E. (1913). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie*, Max Niemeyer Verlag, Halle

Idler, E.L. & Benyamini, Y. (1997). Self-related health and mortality: A review of twenty-seven community studies. *Journal of Health and Behavior*, 38, pp. 21-37, 2150-6000

Jahoda, M. (1958). *Current Concepts of Positive Mental Health*, Basic Books, New York

Jané-Llopis, E. & Braddick, F. (Eds). (2008). *Mental Health in Youth and Education. Consensus paper*, European Communities, Luxembourg, 978-92-79-09526-9

Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, 43, pp. 207-222, 2150-6000

Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, 73, pp. 539–548, 0022-006X

Keyes, C. L. M. (2006). Subjective Well-Being in Mental Health and Human Development Research Worldwide: An Introduction. *Social Indicators Research*, 77, 1, pp. 1-10, 1573-0921

Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. *American Psychologist*, 62, pp. 95–108, 0003-066X

Kilpeläinen, K., Aromaa, A. & the ECHIM project (Eds.) (2008). *European Health Indicators: Development and Initial Implementation. Final Report of the ECHIM project*, National Public Health Institute, 978-951-740-858-5, Helsinki

Koivumaa-Honkanen, H., Honkanen, R., Viinamaki, H., Heikkila, K., Kaprio, J. & Koskenvuo, M. (2001). Life satisfaction and suicide: A 20-year follow-up study. *American Journal of Psychiatry*, 158, pp. 433-439, 1535-7228

Koivumaa-Honkanen, H.T., Honkanen, R., Koskenvuo, M., Viinamaki, H. & Kaprio, J. (2002). Life satisfaction as a predictor of fatal injury in a 20-year follow-up. *Acta Psychiatrica Scandinavica*, 105, pp. 444–450, 1600-0447

Koivumaa-Honkanen, H. T., Kaprio, J., Honkanen, R., Viinamaki, H. & Koskenvuo, M. (2004a). Life satisfaction and depression in a 15-year follow-up of healthy adults. *Social Psychiatry and Psychiatric Epidemiology*, 39, pp. 994-999, 1433-9285

Koivumaa-Honkanen, H.T., Koskenvuo, M., Honkanen, R., Viinamaki, H., Heikkila, K. & Kaprio, J. (2004b). Life dissatisfaction and subsequent work disability in an 11-year follow-up. *Psychological Medicine*, 34, pp. 221-228,

Korkeila, J., Tuomi-Nikula, A., Wahlbeck, K., Lehtinen, V. & Lavikainen J. (2006). Proposal for a harmonised set of mental health indicators, In: *Improving Mental Health Information in Europe*, Lavikainen, Fryers & Lehtinen, pp. 107-116, Stakes and European Union, Helsinki

Kramers, P.G.N. (2003). The ECHI project. Health indicators for the European Community. European Journal of Public Health, 13, pp. 101-106, 1464-360X

Kramers, P.G.N. & the ECHI team (2005). *Public Health indicators for the European Union: Context, selection, definition. Final Report by the ECHI Project Phase II*, National

Institute for Public Health and the Environment, Bilthoven, 31.10.2011, Available from: < http://rivm.openrepository.com/rivm/bitstream/10029/7294/1/271558006.pdf>

Land, K. C. (1975). Social indicators models: An overview, In: *Social Indicator Models*, Land & Spilerman, pp. 5-36, Russell Sage Foundation, New York

Lehtinen, V. Ozamiz, A., Underwood, L. & Weiss, M. (2005). The Intrinsic Value of Mental Health, In: *Promoting Mental Health: Concepts, Emerging Evidence, Practice*, Herrman, Saxena & Moodie, pp. 46-57, World Health Organization, 92 4 156294 3, Geneva

Lippman, L. H. (2007). Indicators and Indices of Child Well-Being: A Brief American History. *Social Indicators Research*, 83, pp. 39-53, 1573-0921

Lippmann, L.H., Moore, K.A., McIntosh, H. (2009). *Positive indicators of child well-being: a conceptual framework, measures and methodological issues. Innocenti Working Paper No. 2009-21*, UNICEF Innocenti Research Centre, 1014-7837, Florence

Maher, I. & Waters, E. (2005). Indicators of Positive Mental Health for Children, In: *Promoting Mental Health: Concepts, Emerging Evidence, Practice*, Herrman, Saxena & Moodie, pp. 159-168, World Health Organization, 92 4 156294 3, Geneva

Maslow, A. H. (1968). *Toward a psychology of being* (2th editon), John Wiley & Son, 0471293091, New York

Mikkelsson, M., Salminen, J. & Kautiainen, H. (1997). Non-specific muskuloskeletal pain in preadolescents: prevalence and 1-year persistence. *Pain*, 73, pp. 29-35, 0304-3959

Morrow, V. & Mayall, B. (2010). Measuring Children's Well-Being: Some Problems and Possibilities, In: *Health Assets in a Global Context*, Morgan, Davies & Ziglio, pp. 145-167, Springer, 978-1-4419-5920-1, New York

Mortimer, J. & Larson, R. (2002). *The Changing Adolescent Experience: Societal trends and the Transition to Adulthood*, Cambridge University Press, Cambridge

Murphy, H.B.M. (1978). The meaning of symptom-checklist scores in mental health surveys: a testing of multiple hypotheses. *Social Science & Medicine*, 12, pp. 67-75, 0277-9536

Neisser, U. (1967). *Cognitive Psychology*, Appleton-Century-Crofts, New York

Nosikov, A. & Gudex, C. (2003). EUROHIS: Developing Common Instruments for Health Surveys. *Biomedical and Health Research*, 57, 1586033220

Palfrey, J. S., Tonniges, T.F., Green, M. & Richmond, J. (2005). Introduction: Addressing the Millennial Morbidity − The Context of Community Pediatrics. *Pediatrics*, 115, pp. 1121-1123, 1098-4275

Parkinson, J. (2007). *Establishing a core set of national, sustainable mental health indicators for adults in Scotland: Final report*, Public Health Adviser and NHS Health Scotland, Edinburgh, 31.10.2011, Available from: http://www.healthscotland.com/uploads/documents/5798-Adult%20mental%20health%20indicators%20-%20final%20report.pdf

Parkinson, J. (2009). *Children and Young People's Mental Health Indicators: Background Briefing*, Public Health Observatory and NHS Health Scotland, Edinburgh, 31.10.2011, Available from: <http://www.healthscotland.com/uploads/documents/9694-C&YP%20Mental%20Health%20Indicators%20Background%20Briefing%20-%20May%202009%20Final.pdf>

Pavot, W. G. & Diener, E. (1993). Review of the Satisfaction with Life Scale. *Psychological Assessment*, 5, pp. 164-172, 1040-3590

Rask, K., Asted-Kurki, P., Paavilainen, E. & Laippala, P. (2003). Adolescent subjective well-being and family dynamics. *Scandinavian Journal of Caring Sciences*, 17, pp. 129–138, 1471-6712

Ravens-Sieberer, U., Gosch, A., Abel, T., Auquier, P., Bellach, B.M., Bruil, J., Dür, W., Power, M., Rajmil, L., & European KIDSCREEN Group (2001). Quality of Life in children and adolescents – a European public health perspective. *Sozial- und Präventivmedizin*, 46, pp. 297-302, 1420-911X

Ravens-Sieberer, U., Gosch, A., Rajmil, L., Erhart, M., Bruil, J., Duer, W., Auquier, P., Power, M., Abel, T., Czemy, L., Mazur, J., Czimbalmos, A., Tountas, Y., Hagquist, C., Kilroe, J. & Kidscreen Group Europe (2005). The KIDSCREEN-52 Quality of life measure for children and adolescents: development and first results from a European survey. *Expert Review in Pharmacoeconomics & Outcomes Research*, 5, pp. 353-364, 1473-7167

Ravens-Sieberer, U., et al. & the European KIDSCREEN Group. (2006). *The KIDSCREEN questionnaires – Quality of life questionnaires for children and adolescents – Handbook*, Pabst Science Publisher, Lengerich

Ravens-Sieberer, U., Wille, N., Erhart, M., Nickel, J. & Richter, M. (2008a). Socioeconomic inequalities in mental health among adolescents in Europe, In: *Social cohesion for mental well-being among adolescents*, pp. 26-42, WHO Regional Office for Europe, 978 92 890 4288 8, Copenhagen

Ravens-Sieberer, U., Gosch, A., Rajmil, L., Erhart, M., Bruil, J., Power, M., Duer, W., Auquier, P., Cloetta, B., Czemy, L., Mazur, J., Czimbalmos, A., Tountas, Y., Hagquist, C., Kilroe, J., & KIDSCREEN Group (2008b). The KIDSCREEN-52 Quality of Life Measure for Children and Adolescents: Psychometric Results from a Cross-Cultural Survey in 13 European Countries. *Value in Health*, 11, 4, pp. 645-658

Ravens-Sieberer, U., Erhart, M., Torsheim, T., Hetland, J., Freeman, J., Danielson, M., Thomas, C. and The HBSC Positive Health Group (2008c). An international scoring system for self-reported health complaints in adolescents. *European Journal of Public Health*, 18, pp. 294-299, 1101-1262

Ravens-Sieberer, U., Erhart, M., Rajmil, L., Herdman, M., Auquier, P., Bruil, J., Power, M., Duer, W., Abel, T., Czemy, L., Mazur, J., Czimbalmos, A., Tountas, Y., Hagquist, C., Kilroe, J. & European KIDSCREEN Group (2010). Reliability, construct and criterion validity of the KIDSCREEN-10 score: a short measure for children and adolescents' well-being and health-related quality of life. *Quality of Life Research*, 19, 10, pp. 1487-1500

Remschmidt, H. & Belfer, M. (2005). Mental health care for children and adolescents worldwide: a review. *World Psychiatry*, 4, 3, pp. 147-153

Rigby, M. J., Köhler, L. I., Blair, M. E. & Metchler, R. (2003). Child Health Indicators for Europe. A priority for a caring society. *European Journal of Public Health*, 1, 1, pp. 38-46, 1464-360X

Rogers, C. (1951). *Client-centered therapy: Its current practice, implications and theory*, Constable, 1-84119-840-4, London

Ryff, C.D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, 57, pp. 1069–1081, 0022-3514

Samdal, O., Nutbeam, D., Wold, B. & Kannas, L. (1998). Achieving health and educational goals through schools: A study of the importance of school climate and students' satisfaction with school. *Health Education Research*, 13, 3, pp. 383-397, 1465-3648

Schwarz, N. & Clore G.L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45, pp. 513–523, 0022-3514

Schwarz, N. & Strack, F. (1999). Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications, In: *Well-Being: The Foundations of Hedonic Psychology*, Kahneman, Diener & Schwarz, pp. 61-84, Russell Sage Foundation, 7 0871544237, New York

Sheldon, E.B. & Moore, W.E. (Eds.) (1968). *Indicators of Social Change: Concepts and Measurements*, Russell Sage Foundation, New York

Shucksmith, J., Spratt, J., Philip, K. & McNaughton, R. (2009). *A critical review of the literature on children and young people's views of the factors that influence their mental health*, NHS Health Scotland, Edinburgh, 31.10.2011, Available from: <http://www.healthscotland.com/uploads/documents/10772-Views%20of%20C&YP%20on%20what%20impacts%20on%20their%20mental%20health%20-%20Final%20report.pdf>

Sigerist, H.E. (1941). *Medicine and Human Welfare*, Yale University Press, New Haven

Starfield, B., Katz, H., Gabriel, A., Livingston, G., Benson, P., Hankin, J., Horn, S., Steinwachs, D. (1984). Morbidity in childhood: a longitudinal view. *New England Journal of Medicine*, 310, pp. 824-829

Suldo, S. M., Riley, K. N. & Shaffer, E. J. (2006). Academic Correlates of Children Adolescents' Life Satisfaction. *School Psychology International*, 27, 5, pp. 567-582, 1461-7374

UN General Assembly (1989). *Convention on the Rights of the Child*

UNICEF (1979). *The State of the World's Children*, Oxford University Press, New York

Wade, T. J. & Vingilis, E. (1999). The development of self-rated health during adolescence: An exploration of inter- and intra-cohort effect. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique*, 90, 2, pp. 90-94

Waterman, A. S. (1993). Two Conceptions of Happiness: Contrasts of Personal Expressiveness (Eudaimonia) and Hedonic Enjoyment. *Journal of Personality and Social Psychology*, 46, 4, pp. 678-691, 0022-3514

Westerhof, G., J. & Keyes, C. L. M. (2010). Mental Illness and Mental Health: The Two Continua Model Across the Lifespan. *Journal of Adult Development*, 17, pp. 110-119, 1573-3440, 1573-3440

Wilkinson, R.B. & Walford, W. (1998). The measurement of adolescent psychological health: One or two dimensions? *Journal of Youth and Adolescence*, 27, pp. 443-455, 1573-6601

World Health Organization (1946). Constitution of the World Health Organization, *Proceedings of International Health Conference*, New York, June-July 1964

World Health Organization (2001). *The world health report 2001 - Mental Health: New Understanding, New Hope*, WHO, 92 4 156201 3, Geneva

World Health Organization (2005). *Promoting mental health: Concepts, emerging evidence, practice*, WHO, 92 4 156294 3, Geneva

World Health Organization (2006). *Addressing the socioeconomic determinants of healthy eating habits and physical activity levels among adolescents: Report from the 2006 HBSC/WHO Forum*, WHO Regional Office for Europe, Copenhagen

Zubrick, S. R. & Kovess-Masfety, V. (2005). Indicators of Mental Health, In: *Promoting Mental Health: Concepts, Emerging Evidence, Practice*, Herrman, Saxena & Moodie, pp. 146-166, World Health Organization, 92 4 156294 3, Geneva

3

Assessing the Outline Uncertainty of Spatial Disease Clusters

Fernando L. P. Oliveira[1], André L. F. Cançado[2],
Luiz H. Duczmal[3] and Anderson R. Duarte[1]
[1]*Departament of Mathematics, Universidade Federal de Ouro Preto*
[2]*Departament of Statistics, Universidade de Brasília*
[3]*Departament of Statistics, Universidade Federal de Minas Gerais*
Brazil

1. Introduction

The spatial analysis of disease incidence is a fundamental tool in public health monitoring (Lawson et al., 1999). Suppose that a geographic study area is divided into administrative areas, with known populations at risk and observed cases of disease within a certain period of time. An interesting question is the possible existence of spatial anomalies in the study area: are there localized regions within the map for which the relative concentration of cases among the population at risk is significantly higher than would be expected if the cases were distributed at random? Such anomalies, known as *spatial clusters*, are inherently difficult to delineate, for several reasons (Cancado et al., 2010; Lawson, 2009). Due to the stochastic nature of the number of observed cases of disease, the uncertainty may be elevated in the disease rate estimation for aggregated area maps, especially for small population areas. Thus the most likely disease cluster produced by any given method for the detection and inference of spatial clusters (like SaTScan (Kulldorff, 1999) or any other irregularly shaped scan) is subject to a lot of variation. If it is found to be statistically significant, what could be said of the external areas adjacent to the cluster? Do we have enough information to exclude them from a health program of prevention?

A criterion was proposed (Goovaerts, 2006) to measure the uncertainty of each area being part of a possible localized anomaly in the map, finding error bounds for the delineation of spatial clusters in maps of areas with known populations and observed number of cases. A given map with the vector of real data (the number of observed cases for each area) was considered as just one of the possible realizations of the random variable vector with an unknown expected number of cases. In this methodology, m Monte Carlo replications were performed, considering that the simulated number of cases for each area is the realization of a random variable with average equal to the observed number of cases of the original map. Then the most likely cluster for each replicated map was detected. Finally, to each area a_i it was assigned the number of simulations that a_i was included in a most likely cluster. If an area belonged to the most likely cluster on all the m replications, it was colored as black;

otherwise, if it never was part of a most likely cluster, then it was colored as white, with intermediate shades of gray in-between. A Bayesian variant along these lines, to detect and represent spatial clusters, was also proposed recently Neill (2011).

Another approach to represent the uncertainty in the delineation of spatial clusters appeared recently (Oliveira et al., 2011), employing a ranking based scheme known as *intensity function*. That procedure uses the circular spatial scan statistic (Kulldorff, 1999) to find the circularly shaped most likely cluster for each replicated map. The corresponding m likelihood values (obtained by means of the m Monte Carlo replications) are ranked. For each area a_i , the maximum likelihood value, obtained among the most likely clusters containing the area a_i, is determined. Finally, the intensity function associated to each areaŠs ranking of its respective likelihood value among the m obtained values is constructed. The latest procedure generally produce less biased results when compared with the two previous schemes.

However, the circular spatial scan has several limitations, which were discussed in the literature (Duczmal et al., 2006; Kulldorff et al., 2006). Particularly, the circular window is not suitable to make the correct delineation of irregularly shaped clusters because it either chooses a proper subset of the true cluster (underestimation) or chooses a large circle containing the cluster as a proper subset (overestimation). One important consequence is the reduction of the power of detection (Duczmal et al., 2006). In order to overcome this limitation, many algorithms were recently proposed to detect irregularly shaped clusters, replacing the circularly shaped window scheme for any strategy of finding irregularly shaped solutions. Usually, the only limitation in shape for those clusters is a connectivity requirement. In this work, we will analyze the utilization of irregularly shaped algorithms for the application of the intensity function (Oliveira et al., 2011), compared to the use of the simple circular scan, which was employed as the standard method. Due to the regular shape of the most likely cluster found, a question was left, at least in part unanswered: do all the areas inside the cluster have the same importance from a practitioner perspective? In this work is proposed an application of the intensity function for irregularly shaped algorithms, thus avoiding a potential problem inherent in the use of the circular spatial scan, which may described as the lack of resolution inside the circular cluster. As a consequence, it may be difficult or impossible to distinguish the relative importance of the areas inside the detected circular cluster. As we shall see, this problem does not occur when using irregularly shaped scans. Besides, the maximum allowed size for the most likely cluster has a large influence in the result of the cluster search (Chen J, 2008).

In this work novel results are presented, applying the multi-objective genetic algorithm scan (Duarte et al., 2010; Duczmal et al., 2008; 2007), adapted for the weighted non-connectivity penalty function (Cancado et al., 2010). Also, by allowing several different maximum sizes for the most likely cluster, the possible anomaly could be identified with greater precision. As will be demonstrated in the following sections, much better delineated cluster maps of the intensity function will be generated, as compared with the previous version using the simpler circular scan. As a consequence, the relative importance of individual regions composing the spatial anomalies may be assessed, and several interesting phenomena related to the geographical distribution of chronic and acute diseases may be visualized.

2. The intensity function

In this section we define a criterion to measure the plausibility of each area being part of a possible localized anomaly in the map. Following Oliveira et al. (2011), instead of finding the most likely cluster in the original map with the observed number of cases for each area, we consider maps where the number of cases are replications of a vector of random variables, whose averages are defined based on the observed number of cases of the original map. We formalize this procedure in the following.

The original map has c_i observed cases in the area a_i, $i = 1, \ldots, K$. Now we construct a Monte Carlo replication distributing randomly the $C = \sum_{i=1}^{K} c_i$ cases among the K areas $a_1, \ldots, a_K$ according to a multinomial distribution where the probability associated to the area a_i is c_i/C. Let $V = (s_1, \ldots, s_K)$ the realization of the multinomial random vector where s_i is the number of simulated cases in the area a_i, $i = 1, \ldots, K$, where $\sum_{i=1}^{K} s_i = C$. The cluster finder algorithm (in our setting we use the circular scan or we use the elliptic scan) now finds the most likely cluster MLC_1 with likelihood ratio value LLR_1. The Monte Carlo procedure above is repeated m times, generating a set of m likelihood ratio values $\{LLR_1, \ldots, LLR_m\}$ corresponding to the most likely clusters $\{MLC_1, \ldots, MLC_m\}$. The likelihood ratio values are sorted in increasing order as $\{LLR_{(1)}, \ldots, LLR_{(m)}\}$ for the corresponding most likely clusters found $\{MLC_{(1)}, \ldots, MLC_{(m)}\}$. We now define the *intensity function* $f : \{1, \ldots, m\} \longrightarrow \mathbb{R}$ by $f(j) = LLR_{(j)}$, $j = 1, \ldots, m$.

For each area a_i, let:

$$q(a_i) = \frac{1}{m} \arg \max_{1 \leq j \leq m, a_i \in MLC_{(j)}} f(j), i = 1, \ldots, K$$

If the area a_i does not belong to any of the sets $MLC_{(1)}, \ldots, MLC_{(m)}$ then we set $q(a_i) = 0$. The value $q(a_i)$ represents the quantile of the highest likelihood ratio among the ranked values of the likelihood ratios of the most likely clusters found in the m Monte Carlo replications, which take into account the variability of the number of cases in each area. In this sense, the value $q(a_i)$ may be interpreted as the relative importance of the area a_i as part of the anomaly of the map, where the value $f(a_i)$ represents the maximum likelihood ratio found for the most likely clusters which contain the area a_i. This concept gives more information about the anomaly than the clear-cut division between cluster and non-cluster areas, as given by the usual process of finding the most likely cluster in the original map. See Oliveira et al. (2011) for further details.

3. Genetic algorithm for spatial cluster finding

3.1 Introduction

Genetic algorithms (GA's) constitute an important class of optimization methods. Its importance comes from the fact that the GA's are robust algorithms, in the sense that they are able to treat a wide variety of problems. While some optimization methods require certain assumptions about the problem to be solved, without which these methods fail, the GA's do not require any assumption of continuity, convexity, differentiability and unimodality. In fact,

the only assumption a GA requires is that the function to be optimized presents a "global trend" that can be captured or learned by the algorithm. Of course, not making any kind of assumption and, consequently, not using these characteristics in favor, GA'a tend to be computationally intensive, so its usage is justified for difficult problems.

When looking for a most likely cluster, one faces a challenging otimization problem: given a set R of n regions in a map, some of which are neighbors, find the connected subset S of R that assumes the highest LLR value. By "connected" we mean that, starting from any region in S there's always a path to any other region of S formed by a chain of neighbors, all of them inside S.

Solving this problem exactly means that we would have to look at all of the 2^n subsets of R, test which ones are connected, evaluate their LLR values and pick up the most likely one. For maps with just a few dozens of regions this problem is already intractable. So we need another strategy to find such optimal solution. GA's showed to be a good alternative for the spatial cluster finding problem (Duczmal et al., 2008; 2007).

3.2 The genetic algorithm

The natural evolution of living beings can be compared to an optimization process. In fact, if individuals who are best adapted survive - in the sense of transmitting their genetic information - while less adapted individuals tend to disappear, it is expected that after a number of generations the population is composed of individuals who are generally better adapted than those of earlier generations. This is also the idea behind a genetic algorithm. It tries to simulate the mechanisms of random variation and selection of adaptive evolution. The mechanisms (or genetic operators) that form the basis of a genetic algorithm are:

- crossover operator, which combines the information of two or more individuals - called parents - generating new individuals - called children;
- mutation operator, which applies a random perturbation to the information of an individual, generating a new one;
- selection operator, which defines the probability of an individual to transmit its genetic information (generate children) based on its adaptation level.

In this context, an individual is a candidate-solution to the optimization problem and a population is a set of individuals. For the spatial cluster detection problem a solution - or individual - is a set of connected regions of the map (the candidate cluster). So, the population is a set of lists, each list being a set of regions that form the solution.

Starting with an initial population the GA forms a sequence of generations. At each iteration it applies the selection, crossover and mutation operators to the current population, generating a new population. The GA used in this work was primarily described in Duczmal et al. (2007) and its biobjective versions were used by Duczmal et al. (2008), Cancado et al. (2010) and Duarte et al. (2010).

3.2.1 Generating the initial population

The initial population is generated by a greedy procedure. Given a map with n regions we generate a population of n individuals, each of which is generated from one region of the map. So, starting with a region, the solution incorporates more regions, choosing at each step to aggregate, among all the regions that are neighbors of some region in the actual solution, the one that makes the LLR value to increase the most when added to the solution. The individual grows until it reaches a maximum size set by the user.

3.2.2 The selection operator

Each solution is evaluated by means of its LLR value and this is the adaptation indicator: higher LLR-valued individuals are more adapted. The selection operator will then give more chances to the more adapted individuals to generate offspring. This is done through a mechanism called binary tournment. For each tournment two individuals are chosen from the current population, each individual having the same probability of being chosed. Then we compare the two solutions and the one with higher LLR value is selected. This procedure is repeated n times, thus producing a set of n selected individuals.

3.2.3 The crossover operator

Now, selected individuals have the chance to trasmit their genetic information to new individuals by generating offspring. Crossover is applied to pairs of parents randomly chosen from the list of selected individuals. The offspring is generated in a way that the children inherit characteristis from both parents. In addition, it is well known that GA's particularly designed for a specific problem perform much better than multiple-purpose generic GA's. Thus, it is highly recommended that operators are designed so that they can take advantage of the intrinsic structure of the problem. For example, in our case we would discard any disconnected cluster candidate because it is an infeasible solution. While a generic crossover operator could, most of the time, generate infeasible clusters, we chose to use a crossover operator that ensures that every generated solution is feasible.

The crossover operator described by Duczmal et al. (2007) presents all these features, being capable of efficiently generating feasible offspring having characteristics of both parents. The operator is implemented in sucha a way that it is only possible to perform a crossover between two parents if they share a nonempty intersection. Once this condition is verified, the offspring is generated. Figure 1 shows an example with two parents (A and B) and the generated offspring 1-5. Note that the offspring constitutes a "path" from one parent to another, with child 1 being more like parent A, while child 5 is almost like parent B. In the middle of the figure we can see parents inside the map with the two intersection regions (in gray).

3.2.4 The mutation operator

Each individual generated by the crossover process has a probability of suffering a mutation. Mutation consistis in introducing a random perturbation in the genetic code of the individual. In our case, the mutation consists of adding to or removing from the cluster a randomly chosen region, provided the cluster's connectivity.

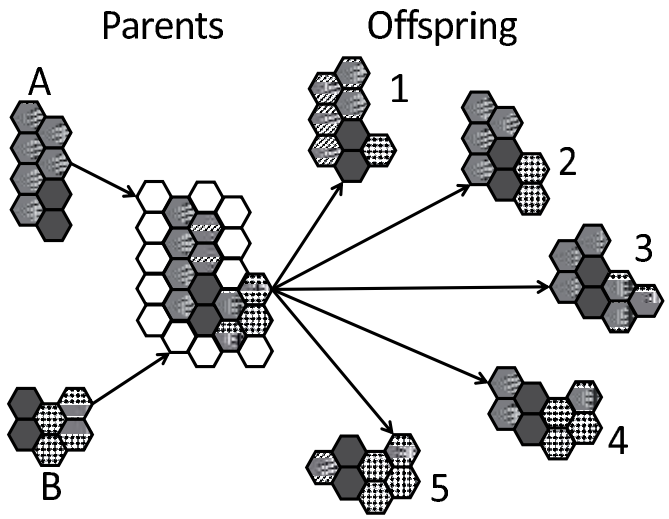


Fig. 1. A splitted vision of parents *A* and *B* (left), parents *A* and *B* inside the map (middle) and offspring (right).

3.3 The biobjective genetic algorithm

Many times one wants to find a solution that simultaneously optimizes two or more functionals. For example, a costumer may want to buy a car which is powerful and cheap. Of course, it is very unlikely that, say, the most powerful is also the cheaper car, because these two criteria are conflicting. Based on these two criteria, a whole set of cars can be of interest for this costumer: powerful (but expensive) cars and cheap (but underpowered) cars. Of course, a costumer (again, based on just these criteria) will reject cars that cost too much and are low powered.

Following the same reasoning, a biobjective GA was proposed (Cancado et al., 2010; Duarte et al., 2010; Duczmal et al., 2008) to deal with the problem of spatial cluster detection. Using the LLR as the only objective to be maximized would lead to geographically meaningless tree-shaped solutions and it is necessary to consider some shape regularity measure, such as geometric compactness (Duczmal et al., 2008) or topological corrections (Cancado et al., 2010; Yiannakoulias et al., 2007). This regularity measure works as a second objective to be maximized. As in the power/price car example, LLR and regularity are conflicting objectives, because high values of LLR are associated to very irregular clusters, while regular solutions tend to

Instead of an optimal solution, a biobjective maximization problem will lead, in general, to a set of optimal solutions, called the Pareto-set. This set is composed by all solutions that are not worse than any other solution in both objectives simulteanously. Such solution is called nondominated solution. Because GA's work with a population of candidate-solutions they can find the Pareto-set in one execution with almost the same effort spent by its mono-objective version. Figure 2 illustrates a set of solutions in the objectives space. Nondominated solutions are indicated by black dots.

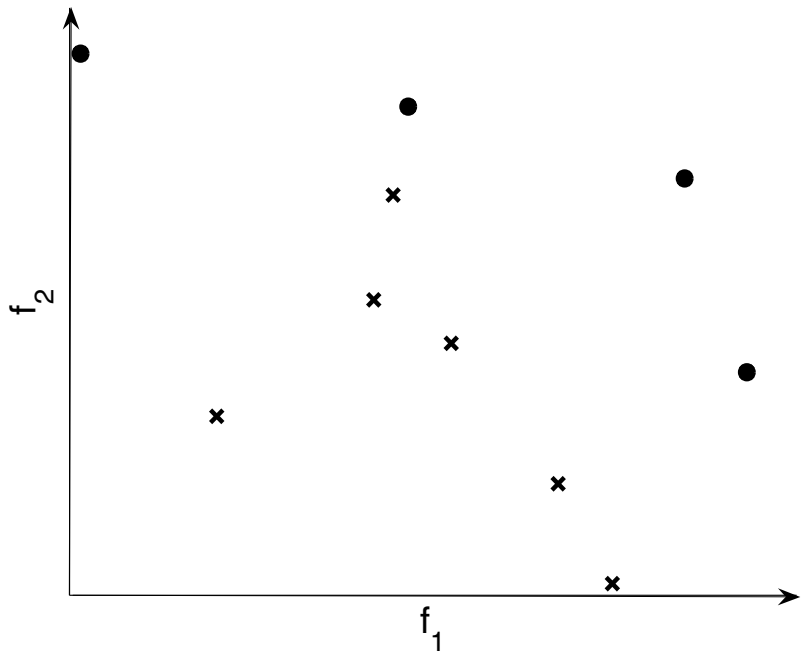


Fig. 2. A set of solutions in the objectives space: dominated solutions ($\times$) and Pareto-set ($\bullet$).

3.4 Inference and the attainment function

Once the most likely cluster is identified, we want to check its significace. This will allow the practitioner to verify if the cluster can be considered a disease outbreak or if the disease cases are randomly spreaded over the map. Since the distribution of LLR under H_0 is not known we must perform a Monte Carlo simulation. For the mono-objective case, the LLR value is calculated for the most likely cluster in each Monte Carlo replication under H_0 and the p-value is computed comparing the value of LLR for the observed data and the empirical distribution obtained through the Monte Carlo procedure.

For the biobjective case, we consider the attainment function (da Fonseca et al., 2001; Fonseca et al., 2005), as also used by Cancado et al. (2010). A single run of the biobjective GA would produce a Pareto-set, defining two distinct regions in the objectives space: points that are dominated by the Pareto-set and points that are not dominated by it. Then, for inference purposes we can consider, for each point of the Pareto-set obtained for the observed data, the proportion times that the point is dominated by the Pareto-sets under H_0. This is exactly the p-value for that point.

3.5 The geometric penalty function

One of the possible penalties that takes in account the cluster geometric shape is the called compactness geometric penalty function. This penalty function introduced in Duczmal et al.

(2006) aims to penalize zones in the map that have very irregular shape. The compactness geometric function $k(z)$ of a zone z is given by the area of z divided by the area of a circle with the same perimeter as the convex hull of z. The compactness geometric function takes values between zero and one, the circle has the most compact shape ($k(z) = 1$). Compactness depends on the shape of the zone, but not on its size. The expression for $k(z)$ is given by:

$$k(z) = \frac{4\pi A(z)}{H(z)^2} \tag{1}$$

where $A(z)$ is the area of the zone z and $H(z)$ the perimeter of the convex hull of z. The compactness penalyzed scan statistic is defined as $max_{z \in Z} k(z).LLR(z)$.

3.6 The non-connectivity penalty function

Yiannakoulias et al. (2007) proposed a greedy algorithm to scan the set Z of all possible zones z. A new penalty function called non-connectivity was proposed. It was based on the ratio of the number of nodes $v(z)$ to the number of edges $e(z)$ of the subgraph associated with the zone z. The non-connectivity penalty was used as a multiplier for the $LLR(z)$. The non-connectivity penalty function of a zone z is defined by:

$$nc(z) = \frac{e(z)}{[3\,(v(z) - 2)]} \tag{2}$$

the expression in the denominator represents the maximum number of edges of a planar graph given its number of vertices. The most penalized zones are the ones whith tree-like associated graphs, meaning that they have a small number of nodes compared with the number of edges. Although there is some similarity between the non-connectivity penalty to the geometric compactness penalty, there is an important difference: the non-connectivity penalty does not rely on the geometric shape of the candidate cluster, which could be an interesting feature when searching for real clusters which are highly irregularly shaped, but present good connectivity properties.

3.7 Evaluation of the candidate solutions

Differently from the previous procedure employing the circular scan, each run of the multiobjective genetic scan produces a set of several non-dominated solutions.

In the circular scan, the scan statistic value for the most likely cluster was assigned to each area of the solution cluster, and later the maximum value of this quantity was obtained for all the executions. However, in the multiobjective procedure, the scan statistic value will be assigned for each component area of each solution cluster of the non-dominated solution set. In the event that a given area belongs to more than one solution cluster, the largest scan statistic value is assigned to the area. The remaining of the process is identical to the usual procedure using the circular scan, obtaining the maximum value of this quantity for all the executions, and building the intensity function as usual.

4. Results and discussion

Epidemiological surveillance is essential to monitoring possible changes in the geographical distribution pattern of both acute and chronic diseases. To illustrate the techniques presented in this chapter, four diseases (dengue fever, tuberculosis, diabetes and hypertension) are analyzed. Those four diseases are currently among the most serious health threats to the Brazilian population. Our studies were concentrated in the Minas Gerais state in southeast Brazil, with 853 municipalities and total population of 19,597,330 (2010 census). For each disease, only the specific population at risk at each municipality was considered. Population data was available at Instituto Brasileiro de Geografia e Estatística (www.ibge.gov.br), and disease data was obtained through DATASUS, the Brazilian Ministry of Health's central data system (www.datasus.gov.br). Dengue fever data was collected by SINAN/MS system from the Brazilian Ministry of Health (www.sinam.org.br). During the period 2007-2010, 349.005 cases were registered, and the population at risk was assumed to be the total population of the 2010 census. Tuberculosis disease cases, using SINAN/MS data, were considered for the 2001-2010 period for the following age groups (years): 15-19, 20-39 and 40-59, making a total of 41,824 cases for a population at risk of 12,892,744. Hypertension data was obtained through the Hiperdia program of Brazilian Ministry of Health from January 2002 to January 2011. Data was available to the following age groups (years): 50-59, 60-69, 70-79 and 80+, with a total population at risk of 4,365,352 individuals and 941,710 cases. Diabetes types 1 and 2 data were also obtained through the Hiperdia program from January 2002 to May 2011. The age groups were: 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 and 80+ years, with 28.039 cases.

Diabetes mellitus and hypertension are considered chronic diseases and their control and treatment depend on the individuals behavior in relation to their lifestyle: healthy eating, physical activity, and weight control. These diseases are responsible for high rates of hospital expenses, so the investment in shares of health promotion and prevention is potentially very cost effective. The importance of dengue in our study lies in the fact that it is an infectious disease and even in regions with previous low incidence rates are subject to outbreaks. This disease is subject to major public health campaigns in Brazil. The report on the epidemiology of dengue published by the Secretariat of Health Surveillance in 2010 indicates Minas Gerais state as one of the critical states in need of stricter monitoring. Hypertension and diabetes are very common chronic diseases, and hypertension is a major public health problem in Brazil. Tuberculosis has become relevant to this study due to its high incidence, and its early diagnosis and effective treatment are of great importance to public health. The biggest challenge for public health professionals has been to promote action to encourage compliance and continuity of care, since many individuals do not join or do not follow the prescribed treatment.

4.1 Real data case studies

In what follows, we present the obtained sets of intensity function maps for dengue fever, tuberculosis, diabetes and hypertension in Minas Gerais state (Figures 3, 4, 5 and 6, respectively). North is up for all the maps. For each disease set we present six maps: (a) the quantiles of population at risk, (b) the quantiles of disease rate and the intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30

and 40 (c, d, e and f respectively). The population at risk was different for each disease in our study.

As can be noted on all four disease sets, the probability that each area belongs to the "'true"' cluster decreases as the maximum cluster size increases from 10 to 40. For instance, in the dengue fever set, the dark brown areas have probability of belonging to the "'true cluster"' greater than 94%, 88%, 83% and 76%, as the maximum cluster size increases from 10, 20, 30 and 40, respectively. It means that the intensity function maps produced with the smaller maximum cluster sizes (10 and 20)indicate inner "'core"' regions within the "'true cluster"'. On the other hand, the intensity function maps produced with the larger maximum cluster sizes (30 and 40)indicate "'borderline"' regions with respect of the "'true cluster"'.

Another important feature is the complexity of the shapes displayed in the sequence of intensity function maps as the maximum cluster size increases.

4.1.1 Dengue fever

In Figure 3c, the maximum size 10 inner core region of dengue fever includes the municipalities arround the state capital Belo Horizonte urban area (population 4 million) in the central part of the state. The maximum sizes 20 and 30 intensity function maps (Figures 3d and 3e respectively) show the anomaly spreading northward following the São Francisco river basin, a region with elevated humidity and high mosquito incidence. Finally, the larger maximum size 40 anomaly (Figure 3f) spreads along the highway joining the cities of Ipatinga, Valadares and Teofilo Otoni in the eastern part of the state.

4.1.2 Tuberculosis

In Figure 4c, the maximum size 10 inner core region of tuberculosis includes the predominantly urban area of Belo Horizonte (in the central part of the state) and two weaker urban regions: (i) the highway joining the cities of Ipatinga, Valadares and Teofilo Otoni in the eastern part of the state, and (ii) the areas surrounding the city of Juiz de Fora, the second largest city of the state in the south. As the maximum cluster size increases (Figures 4d, 4e and 4f), the tuberculosis anomaly is reinforced to include the surrounding municipalities, and also the neighbors of the populous Montes Claros city in the northern part of the state.

4.1.3 Diabetes

In Figure 5c, the maximum size 10 inner core region of diabetes includes the southwest part of the state and the weaker urban region of Valadares city in the east. As the maximum cluster size increases (Figures 5d, 5e and 5f), the diabetes anomaly is reinforced to include the surrounding municipalities.

4.1.4 Hypertension

In Figure 6c, the maximum size 10 inner core region of hypertension includes several scattered regions in the center and mid southeast parts of the state. As the maximum cluster size increases (Figures 6d, 6e and 6f), the hypertension anomaly is reinforced to include the surrounding municipalities.

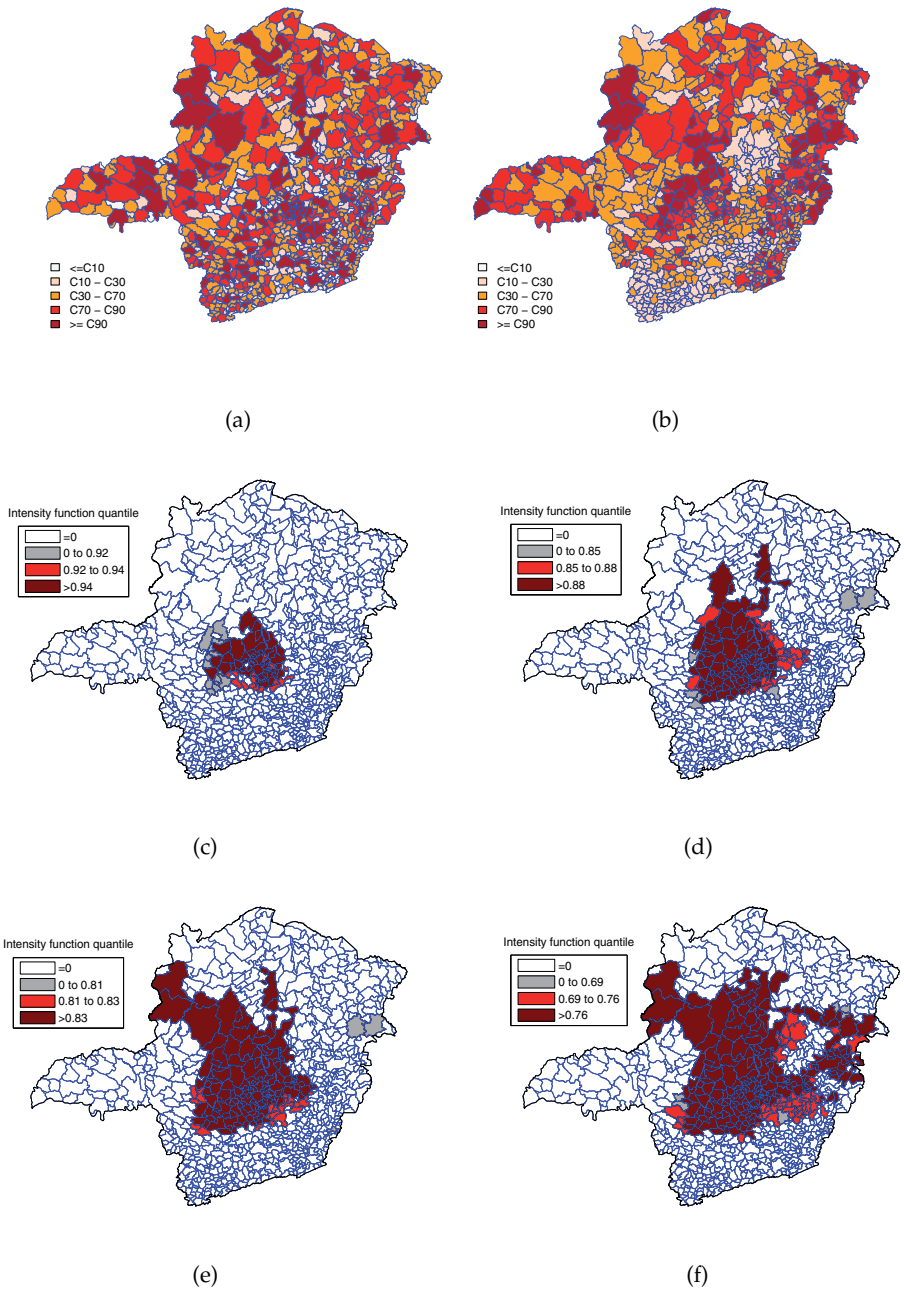


Fig. 3. Population at risk quantiles (a), dengue fever rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

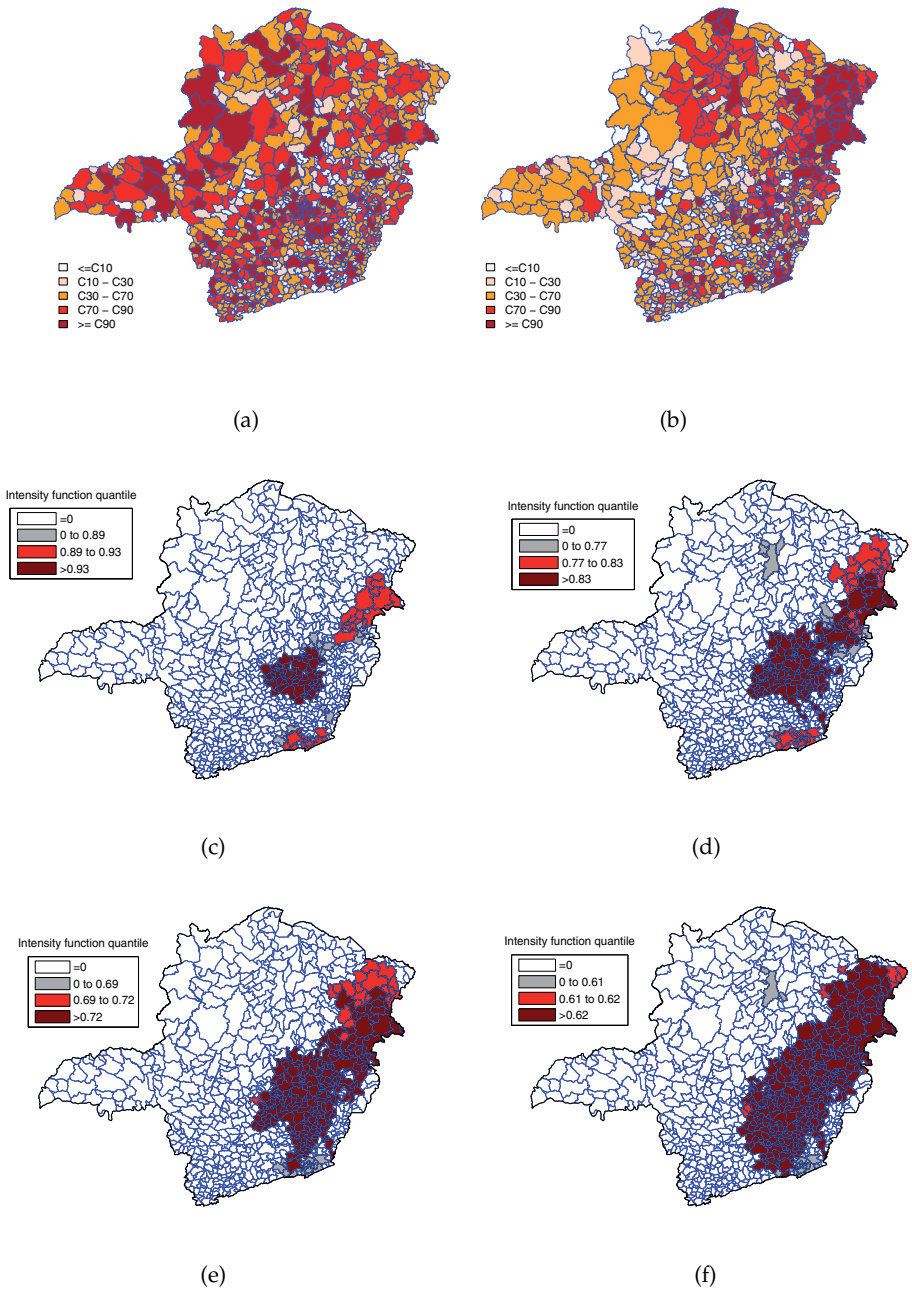


Fig. 4. Population at risk quantiles (a), tuberculosis rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

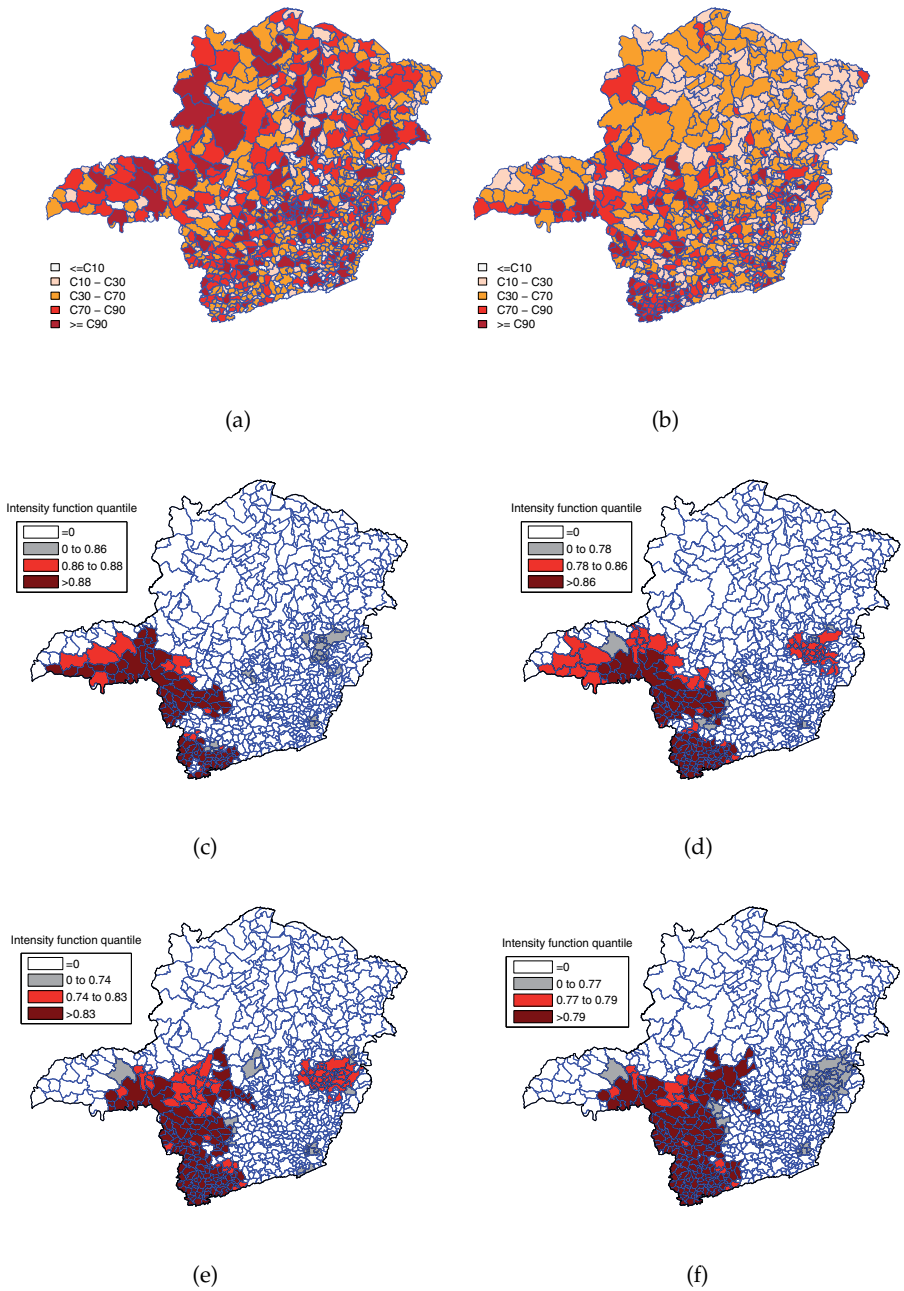


Fig. 5. Population at risk quantiles (a), diabetes rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

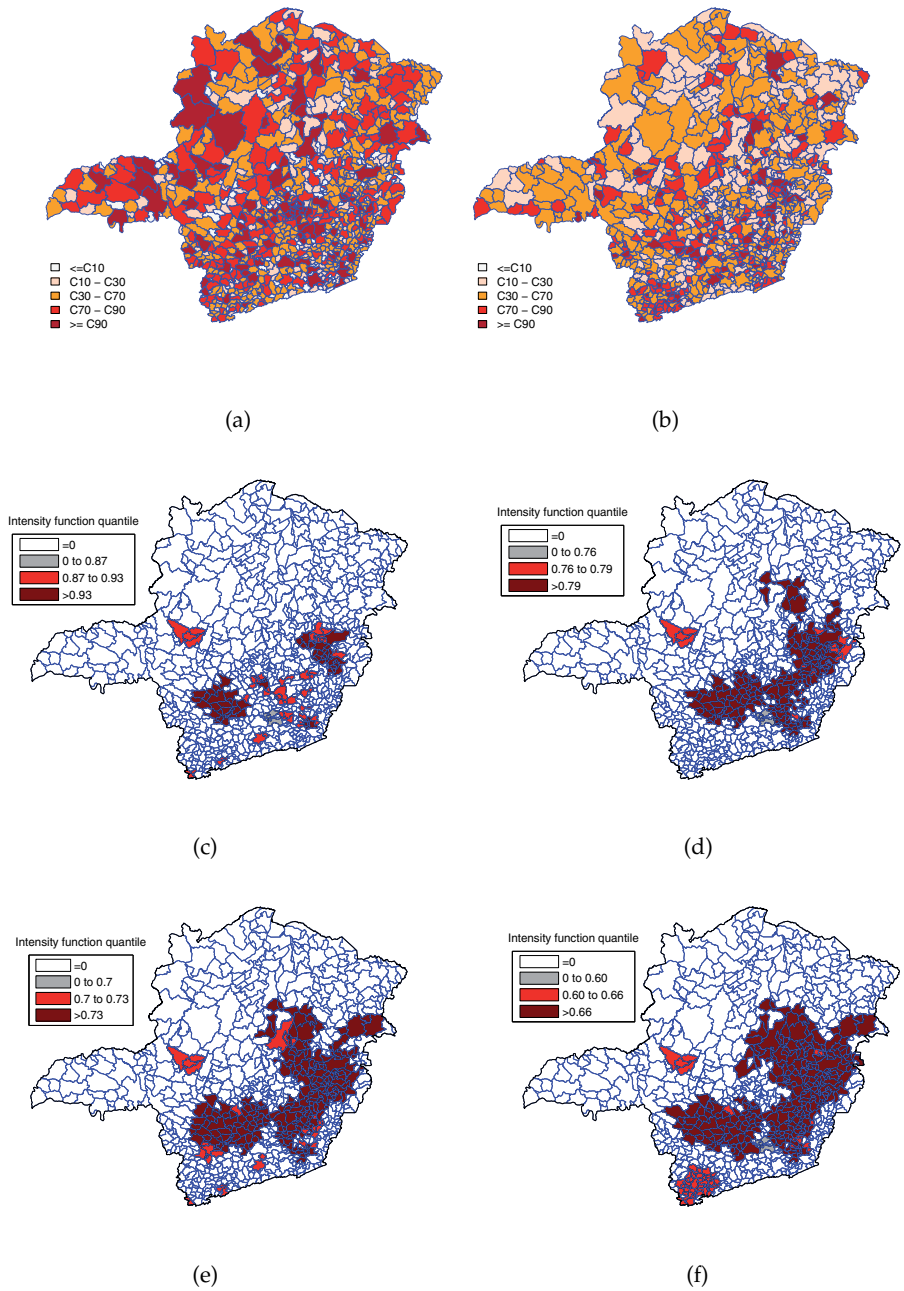


Fig. 6. Population at risk quantiles (a), hypertension rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

5. Conclusion

Our methodology takes into account the variability in the observed number of disease cases on area aggregated maps to nonparametrically infer the uncertainty in the delineation of spatial clusters. A given real data map is regarded as just one possible realization of an unknown random variable vector with expected number of cases. The real data vector of the number of observed cases in each area is used to construct a new vector of expected values of random variables, considering the count of cases as the average of the random variables. This vector is now an estimate of the unknown random variable vector with expected number of cases. Our methodology performs m Monte Carlo replications based on this estimated vector of averages. The most likely cluster of each replicated map is detected and the m corresponding likelihood values obtained in the replications are ranked. For each area we determine the maximum likelihood value among the most likely clusters containing that area. Thus, we obtain the intensity function associated to each area's ranking of their respective likelihood value among the m values. The intensity of each area can be interpreted as the importance of that area in the delineation of the possibly existing anomaly on the map, considering only the initially given information of the observed number of cases. This procedure, based on the empirical distribution, takes into account the intrinsic variability of the observed number of cases, which generally is not considered directly in the existing algorithms used to detect spatial clusters.

In our case studies we could see different situations with respect to the intrinsic variability of the existing spatial anomaly. When the most likely cluster is quite prominent, as in the diabetes case study, the intensity function is such that almost all areas associated with the most likely clusters found in the m replications coincides with those areas composing the most likely cluster detected for the original observed cases. In this situation the geographic anomaly is highly focused. However, in a different scenario, a disease map may present an intrinsically wide variability of data. Many areas near or adjacent to the most likely cluster have values of the intensity function close to the values corresponding to areas of the most likely cluster. In the case study of hypertension, this intrinsic variability produces a map with clearly unrelated areas, but with rather close probability ranking, indicating a situation of multiplicity of clusters, i. e., the most likely cluster is clearly poorly delineated.

In this work we included two new features that extended the original ideas of the previous paper Oliveira et al. (2011). First, instead of the circular scan, we have used an irregularly shaped cluster finder based on a multiobjective genetic algorithm. It allowed a much better delineation of the complex shapes found in the real data clusters. As a consequence, several new phenomena could be distinguished in the spatial distribution of disease, which could not be observed with the simples spatial scan. The second modification was the sequential execution of runs with different sizes for the maximum allowed cluster to composing the intensity function maps. With this modified procedure, instead of only one map, it was obtained a sequence of intensity function maps: as the maximum cluster size increased, larger anomalies of lesser intensity were displayed. This allowed the identification of "'core"' and "'borderline"' regions, with different levels of uncertainty.

The visualization tool developed in this work may serve as a support for the decision making process to prioritize areas of public health intervention, in a more precise manner than provided by ordinary methods of cluster finding.

6. References

Cancado, A. L. F., Duarte, A. R., Duczmal, L., Ferreira, S. J., Fonseca, C. M. & Gontijo, E. C. D. M. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters, *International Journal of Health Geographics* 55(9).

Chen J, Roth RE, N. A. L. E. M. A. (2008). Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of u.s. cervical cancer mortality, *International Journal of Health Geographics* 7(57).

da Fonseca, V. G., Fonseca, C. M. & Hall, A. O. (2001). Inferential performance assessment of stochastic optimisers and the attainment function, *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization, Lecture Notes In Computer Science*, Vol. 1993, Springer-Verlag, Berlin, pp. 213–225.

Duarte, A. R., Duczmal, L. H., Ferreira, S. J. & Cancado, A. L. F. (2010). Internal cohesion and geometric shape of spatial clusters, *Environmental and Ecological Statistics* 17: 203–229.

Duczmal, L., Cancado, A. L. F. & Takahashi, R. H. C. (2008). Delineation of irregularly shaped disease clusters through multiobjective optimization, *Journal of Computational and Graphical Statistics* 17(2): 243–262.

Duczmal, L., Cancado, A. L. F., Takahashi, R. H. C. & Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics, *Computational Statistics and Data Analysis* 52: 43–52. DOI:10.1016/j.csda.2007.01.016.

Duczmal, L., Kulldorff, M. & Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped clusters, *Journal of Computational and Graphical Statistics* 15(2): 428–442.

Fonseca, C. M., da Fonseca, V. G. & Paquete, L. (2005). Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function, *Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization, Lecture Notes In Computer Science*, Vol. 3410, Springer-Verlag, Berlin, pp. 250–264.

Goovaerts, P. (2006). Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using poisson kriging and p-field simulation, *International Journal of Health Geographics* 5(7).

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications, *in* J. Glaz & N. Balakrishnan (eds), *Scan Statistics and Applications*, Springer Netherlands, pp. 303–322.

Kulldorff, M., Huang, L., Pickle, L. & Duczmal, L. (2006). An elliptic spatial scan statistic, *Statistics in Medicine* 25: 3929–3943.

Lawson, A. (2009). *Bayesian Disease mapping*, CRC Press.

Lawson, A., Biggeri, A. & Bohning, D. (1999). *Disease mapping and risk assessment for public health*, John Wiley and Sons, New York.

Neill, D. B. (2011). Fast bayesian scan statistics for multivariate event detection and visualization, *Statistics in Medicine* 30(28): 455–469.

Oliveira, F. L. P., Duczmal, L., Cancado, A. L. F. & Tavares, R. (2011). Nonparametric intensity bounds for the delineation of spatial clusters, *International Journal of Health Geographics* 1(10).

Yiannakoulias, N., Rosychuk, R. J. & Hodgson, J. (2007). Adaptations for finding irregularly shaped disease clusters, *International Journal of Health Geographics* 6(28).

4

Review of Ames Assay Studies of the Urine of Clinical Pathology and Forensic Laboratory Personnel and Other Occupations, such as Oncology Hospitals and Nursing Personnel

Majid Rezaei Basiri[1,5], Mahmoud Ghazi-khansari[2], Hasan Rezazadeh[1], Mohammad Ali Eghbal[1,4*], Iraj Aswadi-kermani[3], H. Hamzeiy[1], Hossein Babaei[1], Ali Reza Mohajjel Naebi[1] and Alireza Partoazar[2]

[1]*Department of Pharmacology and Toxicology in School of Medicine of Tabriz, Iran Medical Sciences University*

[2]*Department of Pharmacology in School of Medicine of Tehran, Iran Medical Sciences University*

[3]*Shahid Ghazi Oncology Research Centre Departments in Tabriz, Iran Medical Sciences University*

[4]*Drug Applied Research Centre of Tabriz*

[5]*Students Researches Committee of Tabriz, Iran Medical Sciences University Iran*

1. Introduction

This chapter we refer to mutagenicity activity in the urine samples of persons who are exposed to carcinogenic materials in their occupations, and so some studies evaluated mutagenicity determination in individuals who worked with and are exposed to active potential mutagenic materials. There are some mutagenic compounds present in workplaces such as among nursing personnel in oncology hospitals, farmers' fields, clinical pathology laboratories, clinical forensic laboratories and pharmacology investigation laboratories etc. Also, clinical forensic laboratory personnel use some dangerous solvents such as chloroform which is mixed with other solvents in solutions of preparations of tank thin layer chromatography. They are also exposed to some mutagenic compounds such as formaldehyde, benzene and some solvents and colour regents. We refer to some of the below mentioned compounds, such as benzene, formaldehyde, paraffin, colour regents, organochlorin, smear fixators and so on. The colour regent might be contained in heavy metal carcinogenic substances which were used for smear colouring in clinical pathology laboratories by technicians and clinical forensic laboratories. According to review of some studies, after the filling out of questionnaire forms by these individuals, urine samples were

* Corresponding Author

collected from candidates in all occupations at the end of the week and they were then held in a refrigerator for the urine extraction process. Special instruments – such as vacuum – and special materials such as MilliQ water and some solvents such as methanol and acetonitrile were used for urine solid phase extraction. Some relationship studies showed that urine extracts were prepared by resins of C18 and amberlite XAD-2 because they are particularly absorbent of the carcinogenic compounds of urine; the importance of the above mentioned resins and the sterilisation of urine extracts will be described in this chapter as well. The personnel who work for a long time in these fields are exposed to these potentially carcinogenic materials. According to the review studies, we will show the mutagenic activity evaluation of the urine of some professional personnel by Ames monitoring. After the preparation of the sterile urine extracts of these personnel, they were exposed to salmonella strains of bacteria for the determination of mutagenic activity. The function of salmonella typhimorium strains such as: TA98,TA100, and the overnight culture of these bacteria show the importance of suspension preparation in the Ames test ratio and all kinds of bacterial environment cultures and the holding of the above mentioned bacteria will be considered in this chapter too. All methods of Ames assay application in these work places will be explained in detail in this chapter. And finally, a risk-benefit assessment and the conditions of healthy work places and individuals' protection will be described as well at the end of this chapter.

2. Mutagenic or carcinogenic material

Some compounds are present in the occupations of clinical pathology and forensic laboratory personnel and other occupations, such as oncology hospitals. Others in this review are: hamatoxilene, eosin and some solvents and colour reagent. Benzene, formaldehyde, paraffin, colour reagents, organochlorin, smear fixators, antineoplastic drugs etc. So the personal of these places might absorb these materials in their work environment through the lungs, the skin, orally etc. These compounds are metabolised in their liver by the P450 enzyme to activate potential mutagenic compounds. Figure.1 shows some these compounds. [5,6,7,8,9,10,12,13,14,15,17,18,19,20,21]

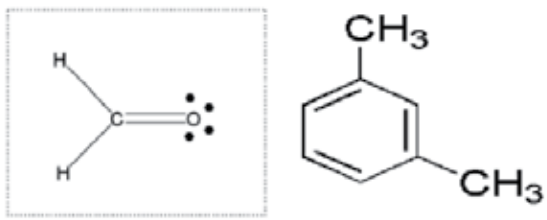


Fig. 1. Mutagenic or Carcinogenic material in laboratory, oxide metabolites of these compounds cause genotoxic effects.

3. Samples and Ames test history

For the mutagenicity activity evaluation of biological fluids in different job environments in previous decades, numerous studies of the founder of mutagenicity have a review up

technique called Ames assay. The first experiments by professor Ames and his co-workers were conducted in the 1960s and 1970s and they were accepted by international agencies, including the American FDA. The test has its own specific conditions because it shows that mutagenicity activities with cost less and the samples of tested results would be expressed in a few days. As well as this, it can simultaneously test up some carcinogenic material over the same time in less than three days. It is held that it also tested for ease of mutagenicity activities which are available for review, of which there may be more in the future. However, in many work environments there are promutagenic and carcinogenic compounds that are present to which workers are exposed. The review of studies of more jobs and environments globally show that using this test for biological fluids – including the urine of personnel – proves potential mutagenic activities. It is also held that the assessment of them in all jobs as well as review by the study process including the preparation of urine samples, extraction of urine mutagenic compounds and the Ames test performed on the samples and biological extracted compounds have similarities processes. [1,2,3,6,12,13,17]

4. Sampling of volunteers and urine extraction

All of the urine samples were collected at the end of the week from the Iranian forensic organisation's laboratories and other relationship laboratories personnel and stored in a refrigerator, although some appropriate methods are accessible for the urine extracts' preparation, but in this review we almost see solid phase extraction methods. Extraction columns were filled with 1 g from one kind of C18 or XAD-2 resins between two layers of cotton, depending on studies' strategies. First, the resins were activated by Milli-Q water and methanol then the volunteer's clear urine was passed through a column by vacuum and the mutagenic substances were concentrated with V/V of methanol/Acetonitrile in tubes. [3,4,6,12,17]

5. Ames assay

In these studies, mutagenicity was followed by the plates' incorporated procedures, as described by Mortelmans and Zeiger (2000), in overnight cultures of TA98 and TA100 salmonella typhymorium tester strains with and without the S9 mix fraction in the forensic organisation's personnel and just TA100 in the clinical laboratory technicians in Iranian laboratories. For this strain, the spontaneous background number of the revertant was approximately 20-50 for TA98 and 75-200 for TA100 salmonella typhymorium tester strains. In this study, Sodium azide was used as a positive control without using the S-9 mix buffer and the 2-amino anthracene as a positive control with the S-9 mix buffer. DMSO and distilled water were considered as negative controls and were used in this study. [1,2,3,6,12,13,17]

6. Ames assay protocol

The first dose/response was determined for all urine extract samples. In our studies, the appropriate dose of urine extracts was 1/100 for exposed bacterial strains and 0.05ml of overnight bacterial culture (TA98 or TA100) was added in 2ml of melting top agar containing trace amounts of biotin and histidine then, and 0.05ml 1/100 of diluted urine extract with DMSO solvent was added in the melting top agar tube as well; after shaking the tube's contents were transferred in Glucose minimum agar plates which were incubated at 37C for 48 hours. The above mentioned protocols were repeated with the addition of 0.5ml

of the S-9mix buffer with the contents of the rat liver p450 enzyme. After the incubation of all plates with and without S-9mix buffer, the colony counts of all the plates were reported for data collection. [1,2,3,4,6,12,13,17]

7. Ames assay on urine samples of hospitals nursings

Most of the anti-neoplastic drugs are cytotoxic as well as nephrotoxic and cause DNA damage. For example, some of them are: doxorubicin, bleomycin, vinblastine, dacarbazine, methotrexate, fluorouracil , prednisone , epirubicin , irinotecan , leucoverin, prednisone, 6-mercaptopurine, procarbazine, lomustine, cisplatin (platinum), etoposide, 6-thioguanine, dexamethasone. According to figure.2, these drugs (such as cycloposphamide, mechlortamine, melphalan, chlorambucil) are nitrogen mustard group and so they have genotoxic effects because they will link to DNA, and they have carcinogenic effects on individuals who are exposed to them in their occupations. They involve a rapid proliferation of normal tissues (bone marrow, hair follicles and the intestinal epithelium). Therefore, most hospital nursing personnel are chronically exposed to anti-neoplastic drugs, especially during the course of giving therapeutic doses to cancer patients. Some studies have shown mutagenic activity in the urine of these nursing staff through an Ames assay. [3,4,6,7,8,9,11,12,13,17]

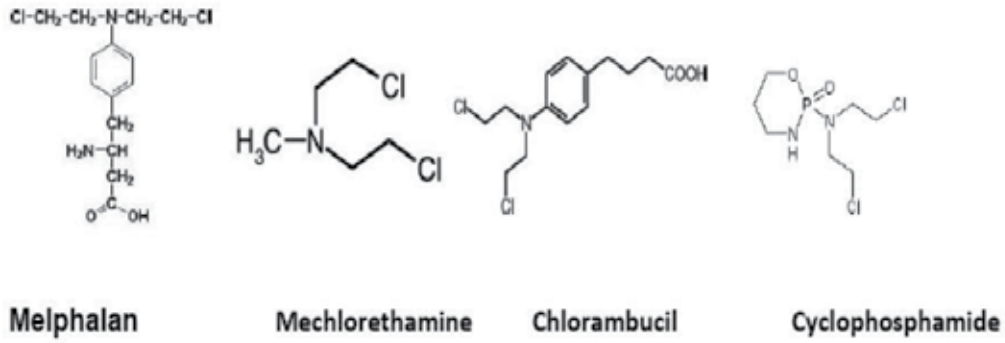


Fig. 2. Some antineoplastic drugs with nitrogen mustard groups.

8. Studies' results

The data was collected from all the occupational personnel and their negative group control urine extracts were exposed to salmonella bacteria tester strains, such as TA98 and TA100 in these review studies; then, according to the below mentioned results and data from the Iranian clinical pathology and medicinal forensic laboratories, they show the colony counts of all the tester strains of salmonella typhymorium bacteria (such as TA 98 and TA100) as being more than 400 for the positive samples of personnel. On the other hand, the colony count of these individuals' urine samples in the bacterial culture shows that these personnel excrete mutagenic compounds in their urine. Each ratio is determined from individuals' urine samples colony counts per negative group of control

persons' urine sample colony counts, so the ratio of all the samples was more than 2 and significant for the description of mutagenic compounds in all occupations. According to figure.3, in these studies we used a 0.05ml DMSO solvent and diluted water for negative control plates with and without using the S-9mix buffer, and we also used sodium azide for positive control plates without using the S-9mix buffer and 2-amino anthracene for positive control plates with the S-9mix buffer.

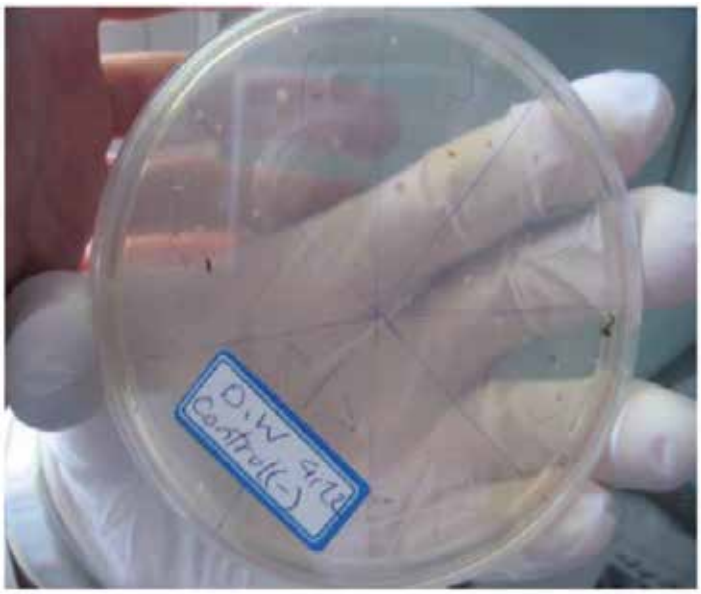


Fig. 3a. Diluted water (D.W) for the negative control colony count plate without using the S-9 mix buffer. Rezai-basiri et al., 2008.

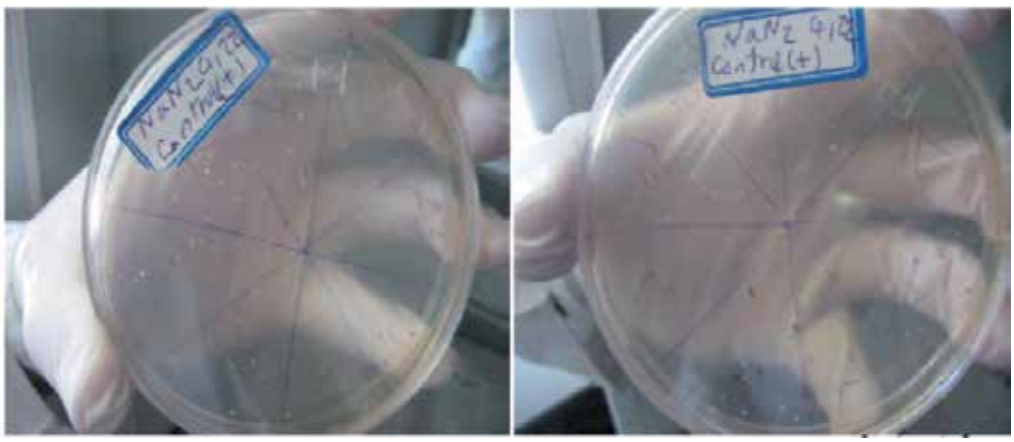


Fig. 3b. Sodium azide (5µg/plate) for the positive control colony count plate (and TA100 Salmonella typhymorium tester strain) without using the S-9 mix buffer. Rezai-basiri et al., 2008.

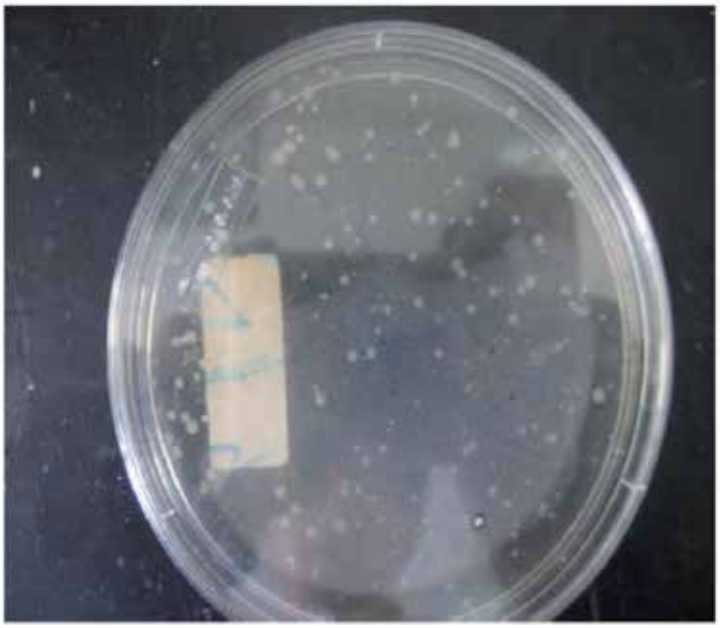


Fig. 3c. DMSO for the positive control colony count plate (and TA100 Salmonella typhymorium tester strain) with using the S-9 mix buffer. Rezai-basiri et al., 2008.

9. Statistical analysis methods

In the Ames assay on the Iranian forensic organisation's laboratory personnel's urines with two bacterial tester strains of TA98, TA100 for colony counts in exposure to urine extract samples, and positive or negative samples the Anova two-way statistical method was used.

10. Evaluation of mutagenic compounds the in urine of medicinal forensic laboratory personnel

Control group	Colony count TA100		Colony countTA98	
	without-S-9mix	with+S-9mix	without-S-9mix	with+S-9mix
Male	86	65	26	25
Male	106	72	23	27
Male	100	77	27	21
Male	85	75	19	24
Male	88	68	22	25
Male	98	65	21	14
Male	95	85	23	29
Female	91	85	20	20
Female	102	86	24	28
Female	98	76	18	24
Average	94.9	75.4	22.2	24.7

Partoazr, et al., 2009.

Table 1. Results of the control group of the mutagenicity assay in the urine of medicinal forensic laboratory personnel with the use of TA98 and TA100 bacteria strains both with and without the S-9mix.

Samples	Colony count TA100		Ratio of TA100	
	with+S-9mix	without-S-9mix	without-S-9mix	with+S-9mix
1*	95	95	1	1.26
2*	95	78	1	1.04
3**	85	74	0.89	0.98

*=The samples belonged to the pathology personnel of the medicinal forensic organisation.
**=The samples belonged to the anatomy personnel of the medicinal forensic organisation. Partoazr, et al., 2009.

Table 2. Positive and doubtful cases. The results of the mutagenicity assay in the urine of medicinal forensic laboratory personnel with the use of TA100 salmonella typhymorium bacteria strains both with and without the S-9mix (cell of rat liver).

Samples	Colony count TA98		Ratio of TA98	
	with+S-9mix	without-S-9mix	without-S-9mix	with+S-9mix
1**	550	55	20	2.2
2*	340	30	15	1.2
3*	41	31	1.86	1.24

**=Positive cases in the personnel of the medicinal forensic organisation.
*=Doubtful cases in the personnel of the medicinal forensic organisation. Partoazr, et al., 2009.

Table 3. Positive and doubtful cases. The results of the mutagenicity assay in the urine of medicinal forensic laboratory personnel with the use of TA98 salmonella typhymorium bacteria strains both with and without the S-9mix (cell of rat liver).

11. Evaluation of mutagenic compound in urine of clinical pathology laboratory technicians

At least four clinical pathology laboratories witnessed the evaluation of the mutagenicity of the urine of technicians by the Ames assay. That the ratio would be more than two was significant for the potential of the mutagenic activity of the urine of technician.

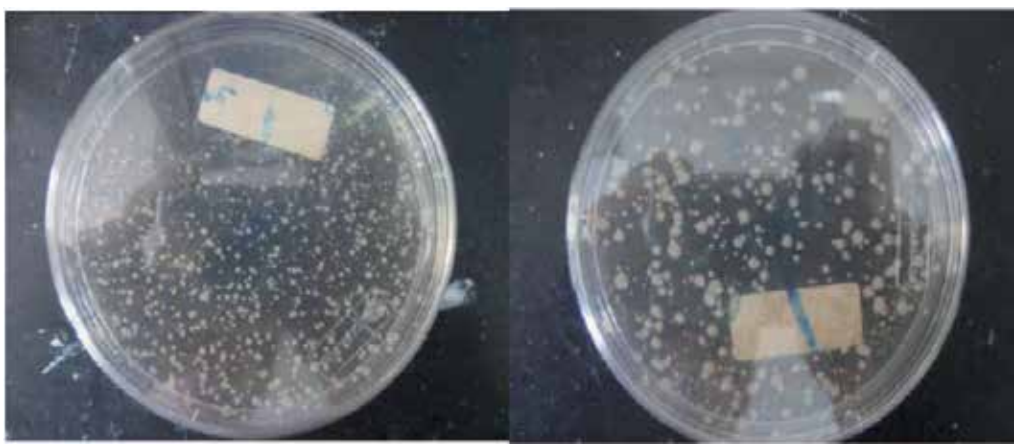


Fig. 4. The plates of positive cases of the TA100 salmonella tester strain cultures at 37C for 48hours for the Ames assay on the urine extracts of the clinical pathology laboratory technicians. Rezai-basiri et al., 2008.

The number of personnel who had greater than two ratios was five people in this study. According to the below mentioned, we see some significant results showing mutagenic excretions in the urine of clinical pathology laboratory technicians:

The ratios without the S-9mix for two of the technicians were 2.01 and 2.0, and the ratio with the S-9mix for these individuals was 2.05 and 2.01. [3,6,12,13,17]

12. Conclusion and recommendation

In this review study, we observed mutagenic activity in the urine extracts of some of the technicians of clinical pathology laboratories and the forensic organisation laboratories. According to these studies, in order to reduce mutagenicity in technicians it was suggested to all of them to use masks, gloves and work under laminar flu and avoid drinking in the laboratories. Considering the contamination of all personnel with mutagenic substances such as colour reagents contained in carcinogenic heavy metals, formaldehyde, benzene, hamatoxilen-Eosin and so on in these laboratories, the observation of the principles of health conditions in clinical pathology laboratories and forensic organisation laboratories is recommended. According to this review study, it is held that the long duration of working and exposure to mutagenic and carcinogenic substances in high risk conditions leads to the excreting of mutagenic compounds in the urine of these individuals and so they should decrease their time spent working over the week. The some studies showed which personnel has used antioxidant compounds, such as thiol group drugs; they had decreased the mutagenic activity in their urine. Also these results were shown to smokers who have used thiol group drugs such as acetyl cystein, had low mutagenic activity in urine samoles. As such, it considers the consumption of vitamins with antioxidant effects which are useful to individuals who are exposed to mutagenic compounds in their occupations. [3,6,12,13,16].

13. Acknowledgments

We are grateful to the below mentioned departments and the research centres of universities and organisation laboratories as well as to the volunteer personnel for their participation in these studies and their professors for the writing of this article:

Department of Pharmacology in the School of Medicine of Tehran/Iran Medical Sciences University.

Department of Pharmacology and Toxicology in the School of Medicine of Tabriz/Iran Medical Sciences University.

Laboratories of Forensic Organization of Tehran/Iran.

Shahid Ghazi Oncology Research Centre Department in Tabriz/Iran Medical Sciences University.

Drug Applied Research Centre of Tabriz/Iran.

Students Researches Committee of Tabriz/Iran Medical Sciences University.

14. References

[1] Mortelmans K, Zeiger E, the Ames Salmonella/Microsome Mutagenicity Assay; Mutat Research, 2000, 29-60.

[2] Zeiger E, Genotoxicity Database, Hand book of carcinogenic potency and genotoxicity database, CRS Press, Boca Raton, FL, 1997, 687-729.
[3] Andre V, Deslandes D, Henry-Amar M, Gauduchon P, Biomonitoring of urine mutagenicity with the Ames test: improvement of the extraction/concentration method, Mutat Res, 2002; 520(1-2): 199-205.
[4] Hyde PM, Evaluation of drug extraction procedures from urine, J Anal Toxicol, 1985, 9(6): 269-72.
[5] Aeschbacher HU, Finot PA, Wolleb U, Interaction of histidine- containing test substances and extraction methods with the Ames mutagenicity test, Mutation Res, 1983; 113(2): 103-16.
[6] Andre V, Lebailly P, Pottier D, Deslandes E, De Meo M, Henry-Amar M, Gauduchon P, Urine mutagenicity of farmers occupationally exposed during a 1-day use of chlorothalonil and insecticides, Int Arch Occup Environ Health, 2003, 76(1): 55-62.
[7] Ahlborg G Jr, Bergstrom B, Hogstedt C, Einisto P, Sorsa M, Urinary screening for potentially genotoxic exposures in a chemical industry, Br J Ind Med, 1985, 42(10): 691-9.
[8] Harrison BR, Developing guidelines for working with antineoplastic drugs, Am J Hosp Pharm, 1981, 38(11): 1686-93.
[9] Zimmerman PF, Larsen RK, Barkley EW, Gallelli JF, Recommendations for the safe handling of injectable antineoplastic drug products, Am J Hosp Pharm, 1981, 38(11): 1693-5.
[10] Anderson RW, Puckett WH Jr, Dana WJ, Nguyen TV, Theiss JC, Matney, TS, Risk of handling injectable antineoplastic agents, Am J Hosp Pharm,1982;39(11):1881-7.
[11] Rezaei-Basiri M, Ghazi-khansari M, Faghih A, Sadeghi M, Lotfalizadeh N, Eghbal M, Mohajell-Nayebi A, Rezazadeh H, Arshad Zadeh M, Screening of Morphine & Codeine in Urine of Opioid Abusers by Rapid and TLC Analysis, Eur J Gen Med, 2010, 7(2): 192-196.
[12] Rezaei-Basiri M, Samini M, Ghazi- khansari M, Rezayat M, Sahebgharani M and Partoazar A, Monitoring Ames Assay on Urine of Clinical Pathology Laboratories Technicians, Journal of Pharmacology and Toxicology, 2008, 3 (3): 230-235.
[13] Partoazar A, Ghazi -Khansari M, Abedi MH, Kaviani M, Norashrafeddin SM, Rezaei-Basiri M,Talebi M, Determining urine sample mutagenicity ratio using Ames test: Tehran forensic medicine laboratory personnel, Tehran University Medical Journal, June 2009, 67(3): 184-189.
[14] Shelef LA and Chin B, Effect of Phenolic Antioxidants on the Mutagenicity of Aflatoxin B1, Applied and Environmental Microbiology, Dec. 1980: 1039-1043.
[15] Cerná M, Pastorková A, Bacterial urinary mutagenicity test for monitoring of exposure to genotoxic compounds: a review, Sep. 2002, 10(3): 124-9.
[16] De Flora S, Camoirano A, Bagnasco M, Bennicelli C, van Zandwijk N, Wigbout G, Qian GS, Zhu YR, Kensler TW, Smokers and urinary genotoxins: implications for selection of cohorts and modulation of endpoints in chemoprevention trials, J Cell Biochem Suppl, 1996, 25: 92-8.
[17] Ames B, Methods for detecting Carcinogens and mutagens with the salmonella/ mammalian micro some mutagenicity test, Mutation Res, 1975, 31: 347-364.

[18] Bartczak AW, Sangaiah R, Ball LM, Warren SH1 and Gold A, Synthesis and bacterial mutagenicity of the cyclopenta oxides of the four cyclopenta-fused of isomers of benzanthracene, Mutagenesis, 2(2), 1987: 101-105.
[19] Varella SD, Rampazo RA, Varanda EA. Urinary mutagenicity in chemical laboratory workers exposed to solvents. J Occup Health, 2008, 50: 415-22.
[20] Siwińska E, Mielzyńska D, Kapka L. Association between urinary 1-hydroxypyrene and genotoxic effects in coke oven workers. Occup Environ Med, 2004, 61: e10.
[21] Chamberlain PL, Brynes SD. The regulatory status of xylazine for use in food-producing animals in the United States. J Vet Pharmacol Ther, 1998, 21: 322-9.
[22] Goodson-Gregg N and De Stasio EA Reinventing the Ames Test as a Quantitative Lab that Connects Classical and Molecular Genetics. Genetics , 2009,181: 21-30.
[23] Ming-Fang Wu, Fu-Chuo Peng, Yung-Liang Chen et al, Evaluation of Genotoxicity of Antrodia cinnamomea in the Ames Test and the In Vitro Chromosomal Aberration Test, in vivo,2011, 419-424.
[24] Meltem Boyacıoğlu, Özlem Çakal Arslan, Hatice Parlak, Muhammet Ali Karaaslan, Mutagenicity of Nonylphenol and Octylphenol Using *Salmonella* Mutation Assay, E.U. Journal of Fisheries & Aquatic Sciences, 2007, 299–302.
[25] Alva Biran , Rami Pedahzur, Sebastian Buchinger, Georg Reifferscheid ,and Shimshon Belkin, Genetically Engineered Bacteria for Genotoxicity Assessment, Hdb Env Chem ,2009,161–186.
[26] Jadwiga Marczewska, Ewa Karwicka et al, Assessment of Cytotoxic and Genotoxic Activity of Alcohol Extract of Polyscias Filicifolia Shoot, Leaf, Cell Biomass of Suspension Culture and Saponin Fraction, Acta Poloniae Pharmaceutica,2011, 703-710.

5

Old Obstacles on New Horizons: The Challenge of Implementing Gene X Environment Discoveries in Schizophrenia Research

Conrad Iyegbe, Gemma Modinos and Margarita Rivera Sanchez
Institute of Psychiatry, Kings College London, UK

1. Introduction

Genetics and Social Sciences are divergent disciplines for whom it is customary to compete to explain the greater part of Schizophrenia risk [1]. These days, a convincing case can be made for the prospective public health value of either discipline [2,3]. However the practical implementation of such knowledge continues to prove challenging for either field alone: From a genetic perspective, progress was traditionally hindered by the inconsistent nature of discoveries made in the pre-GWAS (Genome-wide Association Study) era. It is now held back by the fact that the heritability attributed to this disorder remains largely impermeable to GWAS and other genomic approaches.

Socio-environmental research, on the other hand, has not progressed to the point of being able to pinpoint the precise origins of the high attributable risk fractions repeatedly encountered [2].

However ongoing progress on two fronts is fuelling hopes that a successful marriage of the two fields will benefit both the rate and the integrity of new discoveries, so that clinical interventions can eventually be targeted to patient sub-groups on the basis of their combined genetic and environmental risk profile:

i. Firstly, the credibility of Schizophrenia genetics is benefiting from a recent upswing in the generation of verifiable new findings. This has led to a palpable mood change within the psychiatric genetics community [4].
ii. Secondly, it is anticipated that social science research will benefit from an unprecedented program of investment that will stimulate the emergence of newer methodologies designed to improve the resolution with which social risk factors are measured [5,6].

There are high hopes that the formal integration of these two fields will help to invigorate the search for tailored clinical interventions, whether they be therapeutic or prophylactic in nature. Thus it seems an opportune time to consider the potential obstacles that lie ahead for Schizophrenia research in the newly revitalised era of translational research. We do this by

taking a fresh look through a retrospective lens, at the historical stumbling blocks for the GxE field. We discuss some of the new opportunities (horizons) at the disposal of GxE researchers designed to circumvent them. Some of these hail from recent advances in biobanking, meanwhile new bioinformatic initiatives are helping to transform electronic clinical databases into similarly powerful research tools.

We also highlight the potential pitfalls of an over-regulated clinical trial environment and the detrimental consequences this may eventually have on the pipeline for new drugs. Currently there are fears that an over-burdensome European regulatory legislature is responsible for the recent efflux of companies away from the European clinical trial market. This may create an unwanted bottleneck (or worse still, a precipice) within the new and fully-functional formal framework designed to shepherd only the most robust GxE discoveries into the clinic. We begin this chapter with a brief review of some important concepts central to a discussion on Gene-Environment inter-dependency.

2. The enigma surrounding heritability

Heritability is defined as the proportion of phenotypic variance due to genetic variance. The concept of Schizophrenia as a heritable disorder was once considered to be controversial, though this is no longer the case. From a scientific perspective it is well worth knowing beforehand that a phenotype of potential interest is heritable enough to merit the effort of dissecting genetically. Thus, establishing that this is the case, is a prerequisite first step in genetic research.

Formal estimates of heritability can be obtained through a number of different methods. The archetypal approach uses twins [7]. Twin studies suggest that susceptibility to Schizophrenia is predominantly a genetic phenomenon that accounts for 65-80% of overall risk [7,8]. But that upper estimate is likely to understate the true importance of the environment. Even highly penetrant genetic risk factors (such as a syndromic deletion on chromosome 22q11), are not always sufficient to elicit Schizophrenia on their own [9]. This is confirmed by the fact that pathogenic genetic anomalies are often harboured by asymptomatic controls, as well as cases [10]. This suggests that the underlying risk conferred is heavily mitigated by the environment and other background genetic modifiers of main effects.

Heritability studies estimate that the environmental contribution to Schizophrenia is between 15-35% of the phenotypic variance. The issue of which science explains the greater part of risk is contentious; social science research bases its own claims of dominance on larger explained effects, and also recent calculations which suggest that the burden of cases occurring in the general population could be averted through social interventions [2]. In truth, methodological biases in both fields mean attempting to delineate between the effects of genes and environment is a somewhat arbitrary exercise. This is because classical approaches to heritability estimation do not automatically factor-in the dependency which may occur between genes and environment. Meanwhile, one all-important confounder not accounted for by the social risk liability models of Kirkbride et al [2], is a family history of psychiatric disorder, (a proxy for genetic influence). It is important to keep in mind that these methodological limitations mean that a disorder caused by GxE will be attributed to Genes in a twin study and Environment in an epidemiological study.

This ambiguity probably explains why heritability estimates for Schizophrenia have historically been so variable; the value of each estimate is affected by parameters defined by the population under study, and also the degree to which characteristics such as gender, age, and environment exposure profile have been averaged-over [11]. Therefore it comes as no surprise that genetic epidemiology research in Psychiatry is fast becoming preoccupied with redefining heritability itself [11]. Some principle areas of interest emerging from such work include;

- the stability of heritability over time
- genetic determination of sensitivity to exposure

Twin studies and other methods impose a fixed-point approximation of heritability. But this fails to adequately capture the inherent mobility of heritability over time. Evidence for this drift comes from longitudinal studies of both Substance Misuse and Depression. These demonstrate a tendency for heritability to increase across the developmental period between adolescence and adulthood [12], and also with later stages of decline [13]. These studies show that the initiation of cannabis use is predominantly an environmental phenomenon, although genetic influences become increasingly important as the level of usage progresses towards substance abuse and drug dependency [14]. In extreme scenarios within polygenic disorders, heritability may reach a higher level during earlier neuro-developmental stages. Such cases tend to result in earlier onset. For Schizophrenia, the earliest cases are known to occur during childhood [15].

3. Multifactorial risk factors for Schizophrenia

Table 1 lists some of the important exposures known to affect the risk of Schizophrenia. The main origins are social, socio-economic and neuro-developmental. As well as being very common many of these risk factors are associated with large effects. Odds ratios (ORs), reflect the odds of exposure to a risk factor in cases relative to controls (expressed as a fold-difference).

Table 1. Environmental Risk factors for Schizophrenia - a non-exhaustive list

Context	Environmental Risk Factor	Recent Review	Recent meta-analysis
Social	Urban-Rural dwelling	March et al, 2008 [16]	-
	Social Context - Neighbourhood effects	-	-
	Social Discrimination-Discrimination	-	Allardyce et al, 2005 [17]
	Migration	Cantor-Graae and Selten, 2005 [18]	DeAlberto et al, 2010 [19]
	Cannabis Use	Henquet et al, 2008 [20]	Arsenault et al, 2002 [21]; Henquet et al, 2005 [22]; Moore et al, 2007 [23]
Familial	Childhood Trauma	Morgan and Fisher, 2007 [24]	-
	Advancing Paternal age	Miller et al, 2010 [25]	Miller et al 2011 [26]
	Seasonal birth	Davies et al, 2003 [27]	Davies et al, 2003 [27]
Neurodevelopmental	Birth defects/Obstetric complications	-	-
	Seasonal birth	-	-
	Vitamin D	-	-
Economic	Developed vs Developing Country	-	Saha et al, 2005 [25]
	Socio-Economic status	Cohen et al, 2008 (22)	-
		-	-
Other	Gender	-	Aleman et al, 2003 [23]
		-	McGrath et al, 2004 [24]
		-	-

The typical effect range of the risk factors shown in table 1 typically range from 1.5 to 11. In contrast, common genetic risk factors for Schizophrenia are much smaller, typically with Odds ratios of between 1.1 – 1.4. See table 2 for a summary of genetic risk factors for Schizophrenia deriving from large-scale (genome-wide) genetic studies.

Table 2. Genetic Risk factors for Schizophrenia

Chromosome	Gene/Region	Symbol	Full Gene Name	GWAS Significance threshold	Reference
1	1q21.1 deletion			Too rare to compute	28-36
	1q21.1	*BCL9*	B-cell CLL/lymphoma 9	Strongly suggestive	37
	1p21.3	*MIR137* (intron 3 of miRNA transcript)		Significant	4
	1q24			Significant	38
	1q32.2	*PLXNA2*	plexin A2	Strongly suggestive	39
2	2p16.1	*VRK2*	vaccinia related kinase 2	Significant	40,41
	2p16.3 deletion	*NRXN1*	neurexin 1	Too rare to compute	28,31-35,42-44
	2p22.2	*SULT6B1*	sulfotransferase family, cytosolic, 6B, member 1	Strongly suggestive	45
	2q32.1	*ZNF804A*	zinc finger protein 804A	Strongly suggestive	46-48
	2q32.3	PCGEM1(non-coding RNA transcript)	(prostate-specific transcript 1	Significant	4
	2q33.3-q34	*ERBB4*	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	Strongly suggestive	49
	2q34-q35	*ACSL3-KCNE4*	acyl-CoA synthetase long-chain family member 3	Significant	50
	2q37	*CENTG2/AGAP1*	ArfGAP with GTPase domain, ankyrin repeat and PH domain 1	Strongly suggestive	49
	2q37.1	*UGT1A1-HJURP* (intergenic)	Holliday junction recognition protein	Significant	50
	2q37.3	*AK573765-TWIST2* (intergenic)	twist homolog 2 (Drosophila)	Significant	50
	2q37.3	*LRRFIP1*	leucine rich repeat (in FLII) interacting protein 1	Significant	50
	3p21.1	*PBRM1*	Polybromo 1	Strongly suggestive	51
3	3q21-q23	*RELN*	reelin	Strongly suggestive	52
	3q21-q23	*RBP1*	retinol binding protein 1, cellular	Strongly suggestive	53
	3q39 deletion			Too rare to compute	28,32,34,35,43, 54
5	5q14.1	*CMYA5*	cardiomyopathy associated 5	Strongly suggestive	55
6	6p21	*ZKSCAN4*	zinc finger with KRAB and SCAN domains 4	Significant	56
	6p21.3	*NOTCH4*	notch 4	Strongly suggestive	45,57
	6p22.1	MHC region		Significant	49,57,58
	6p22.1	*NKAPL*	NFKB activating protein-like	Significant	56
	6p22.1	*PGBD1*	piggyBac transposable element derived 1	Significant	56,57
	6q21-qter	*LOC645434-NMBR* (intergenic)	neuromedin B receptor	Significant	50
	6q23.2	*AHI1*	Abelson helper integration site 1	Strongly suggestive	59

Chromosome	Gene/Region	Symbol	Full Gene Name	GWAS Significance threshold	Reference
7	7q11.23-q21.3	*PCLO*	piccolo (presynaptic cytomatrix protein)	Strongly suggestive	60
	7q36.3 duplication	*VIPR2*	vasoactive intestinal peptide receptor 2	Too rare to compute	28,32,53,61,62
8	8p12			Significant	38
	8p21-p12	*NRG1*	neuregulin 1	Strongly suggestive	49
	8q21	*MMP16*	matrix metallopeptidase 16	Significant	4
	8p23.2	*CSMD1*	CUB and Sushi multiple domains 1	Significant	4
11	11q24.2	*NRGN*	neurogranin (protein kinase C substrate, RC3)	Significant	41,57
12	12p13.3	*CACNA1C*	calcium channel, voltage-dependent, L type, alpha 1C subunit	Suggestive	63
15	15q13.2 deletion			Too rare to compute	28,32,34
16	16p11.2 duplication			Too rare to compute	28,32,35
	16p13.11 duplication			Too rare to compute	28,30,31,34
17	17q12 deletion			Too rare to compute	32,34
18	18q21.1	*TCF4*	transcription factor 4	Significant	41,57
22	22q11.21 deletion			Too rare to compute	28,31,32,34,36, 54
X	Xp22.3 & Yp13.3	*IL3RA*	interleukin 3 receptor, alpha (low affinity)	Strongly suggestive	64
	Xp22.32 & Yp11.3	*CSF2RA*	colony stimulating factor 2 receptor, alpha	Strongly suggestive	64

Table 2. CNVs (Copy Number Variants) are sub-microscopic deletions and duplications of DNA (typically greater than 100kb in size). SNP (Single Nucleotide Polymorphism) refers to a single subnit (base) change in the DNA sequence. *CACNA1C, ZNF804A, NRGN, MHC* and *PBRM1*, all overlap with Bipolar Disorder. *Genome-wide significant = $P<5x10^{-8}$; Strong significance is defined as a *P* value of between $5x10^{-4}$ and $5x10^{-8}$. **Notes the pre-existence of this gene as a commonly-researched candidate in Schizophrenia research prior to GWAS. Note within table 2 the high occurence of findings validated by more than one study. This is particularly obvious for CNVs, but is also evident for SNP variants, including those not reaching overall significance.

The effects of Environmental risk factors are on a par with those of the structural variants [9], catalogued in Table 2, but the latter occur much too infrequently to explain the fact that Schizophrenia is common mental disorder, affecting 1% of the global population. In fact, the molecular modalities identified so far for Schizophrenia (namely copy number and common variation) currently account for no more than 3% of the total phenotypic variance of Schizophrenia [65].

The discrepancy between theoretical and observed heritability estimates has led many to speculate on possible reasons why the 'missing' component is so elusive [66]. The possibilities span a wide array of plausible theories, most of which are based on the premise that the additive component of heritability is probably exaggerated. eg [67].

4. The fundamental models of gene-environment dependency

Nowadays, it is possible to examine the theory that heritability has been overstated, by testing the significance of the difference in heritability between exposure and non-exposure twin models [11]. This is an appropriate way to empirically test the dependency between genes and environment.

The risk factors shown in table 1, while very common, are mitigated by the genetic make-up of the individual, such that the overall effect on risk is relatively small. At a population level this means that only a small proportion of those encountering these exposures will ever go on to develop clinical symptoms. This example of inter-dependency is known as Gene-Environment interaction. Cumulatively it may have a large impact on psychosis risk at the population level.

4.1 G-E correlation (rGE)

Analytically, GxE is difficult to distinguish from gene-environment correlation (rGE), a phenomena whereby exposure to exogenous risk factors is encoded within the DNA of the individual. rGE represents the social manifestation of one's genetic heritage, and its influence on subsequent lifestyle choices. If not properly accounted for, rGE can quietly confound the apparent interaction between genes and the environment.

There are many behavioural examples of this phenomenon in the psychiatric literature (reviewed in [68]). For instance, genes can have an indirect influence on adolescent substance misuse, through a mechanism in which genes drive the selection of friends who facilitate this behaviour. In this example, peer-group choice can be redefined as a lifestyle trait with a strong genetic component [69]. An equally compelling case can be made for rGE in Depression, as there is an indication that genetic susceptibility to Depression may also partly reflect a person's tendency to experience stressful experiences, such as interpersonal and romantic difficulties [70].

The evidence used to discuss G-E dependency in the context of Schizophrenia is drawn almost exclusively from the cannabis literature, as it is one of the most commonly investigated risk factors in GxE research. Its popularity probably reflects the relative ease with which data on this exposure may be obtained and verified, with good sensitivity and specificity. This makes it comparatively easy to derive a fairly accurate profile of exposure using retrospective assessments [71,72].

While there is little in the way of direct experimental evidence to support the occurrence of rGE in Schizophrenia, it would be surprising if Schizophrenia were shown to be completely devoid of the phenomenon, given its demonstration in other areas of behavioural research [68]. Only one study has purported to show evidence of the rGE mechanism in Schizophrenia [73]. Meanwhile the evidence that contradicts this finding has withstood the many different experimental designs applied to re-address the same question eg. [21,74,75]. The most recent of these used a case-control design [75], and also included a comparison of lifetime cannabis consumption between the siblings of cases (who have a higher genetic propensity for Schizophrenia) and healthy controls. It found

no difference between these two groups and thus does not find support a role for Schizophrenia genes in the initiation of cannabis use.

4.2 G-E interaction (GxE)

Interaction is a more solidly supported mechanism of G-E dependency in Schizophrenia, whose influence clearly extends to cannabis use. For example, early studies have suggested that familial (presumably genetic) influences on SZ risk also augment the psychotogenetic effects of this drug [76]. Another study finds that the same level of familial liability is reached among cases of cannabis-induced psychosis, as that found among Schizophrenia patients; a strong indication that the enhanced responsiveness to cannabis in these hospitalised users is enabled by Schizophrenia genes [77]. Cannabis use can thus be said to advance the genetic risk of Schizophrenia onset. The same can be said of urbanicity [78] and prenatal exposure to infection [79], but seemingly not of obstetric complications [80].

One drawback of the *familial liability* study design is that genetic and environmental effects cannot so easily be discerned within the construct of 'familiality', which is inferred as being predominantly genetic in origin, but which also incorporates an element of shared environmental risk. The adoption study design therefore, is a convenient way to disentangle the components of this construct, by allowing the genetic component to be assessed in isolation of shared environmental influences.

The adoption study design has been widely used for this purpose in Schizophrenia research. A recent exemplar for the approach investigated psychosis in 13,000 entrants on the Swedish National Adoption Register. Using an empirical approach, the study confirmed the relevance of early life parental employment status, parental separation and housing status to underlying Schizophrenia risk. Importantly this occurred both dependently and independently of underlying genetic liability. The synergism between genes and environment was many times greater than either additive or multiplicative risk thresholds, indicating a strong interaction. The findings were later validated in 26,000 individuals derived from the general population [81].

5. Candidate gene studies of gene-environment interaction

Currently, Gene-Environment interaction is one of a few areas of genetic research in which the candidate-gene design has had the upper hand over the more systematic approach represented by Genome-wide Association studies (GWAS). A favoured approach uses biological plausibility to guide the formulation of coherent hypotheses [82]. This strategy has several high profile discoveries to its credit. Table 3 Lists the GxE studies performed to date in psychosis and summarises the individual outcome of each. Heterogeneity among hypotheses and methodological approaches precludes a more formal assessment of current experimental evidence (ie. by meta-analysis).

Universal acknowledgement of the GxE concept in Schizphrenia alone [78-81,83-86] tends to suggest that its pervasiveness across psychiatry should be on a par with the rest of nature. However the paradox of GxE in Psychiatry is that though generally acknowledged, the interactions themselves are proving difficult to individually identify.

Table 3. Studies investigating interactions between candidate susceptibility genes and candidate environmental pathogens in relation to psychosis.

Author	Sample size	Candidate G	Candidate E	Outcome Variable	Results	Statistics
Zammit et al 2011	2630 HC general population	*COMT* Val(158)Met	Cannabis	Self-reported psychotic experiences at age 16	No interaction	p=0.304-0.981 OR=0.83-1.10
vanWinkel et al 2011	810 SZ, 740 siblings, 419 HC	152 SNPs in 42 candidate genes	Cannabis	Psychotic disorder	Interaction with AKT1 rs2494732 only in cases	p=0.007
Ho et al 2011	235 SZ	12 tag SNPs in *CB1/CNR* gene	Cannabis	Brain Volume and Neurocognition in SZ	- Brain Volume: rs12720071-G-allele carriers with marijuana misuse had the smallest meanparietal WM volumes. rs7766029-C/C associated with small temporal and parietal WMvolumes. rs9450898-C/Cassociated with small frontal and parietalWMvolumes. - Neurocognition: CNR1 rs12720071-G-allele carriers with marijuana misuse had the worst problem solving test performance.	p<0.05 p≤0.05 p≤0.05 mean z=−1.78
Decoster et al 2011	585 SZ	*BDNF* Val(66)Met	Cannabis	Psychotic disorder (Age of onset)	- No BDNFx Cannabis interaction. - Significant BDNFx Cannabis x Sex interaction (females).	p=0.420; $\chi^2(1)=0.65$ p=0.023; $\chi^2(1)=5.15$
Kantrowitz et al 2009	92 SZ (33 Caucasian, 46 African-American)	*COMT* Val(158)Met	Cannabis	Adolescent cannabis use	No association cannabis use-COMT genotypes (African-American/Caucasians)	p=0.23/0.49; $\chi^2(2)=2.9/1.4$
Henquet et al 2009	31 psychotic disorder, 25 HC	*COMT* Val(158)Met	Cannabis (ESM)	Psychotic symptoms (hallucinations, delusions) in daily life (ESM)	Cannabis increased hallucinations in Val/Val carriers with high levels of psychometric psychosis liability.	p<0.001; β=0.78
Zammit et al 2007	750 SZ, 688 HC	*CNR1, CHRNA7, COMT* Val(158)Met	Cannabis, Tobacco	Psychotic disorder	No interaction with CNR1 or COMT genotypes.	p>0.05; OR=0.83-0.98
Henquet et al 2006	30 psychotic disorder, 12 relatives, 32 HC	*COMT* Val(158)Met	Cannabis	D-9-THC-induced psychotic experiences	Condition x Val/Val x Psychometric psychosis interaction.	p=0.003; $\chi^2(1)=8.86$
Caspi et al 2005	803 HC general population	*COMT* Val(158)Met	Cannabis	Schizophreniform disorder	Cannabis x Val/Val x Schizophreniform disorder interaction.	p=0.025; OR=10.9
Alemany et al 2011	533 HC general population	*BDNF* Val(66)Met	Childhood abuse and neglect	Positive and negative psychotic-like experiences	BDNF (Met/-) x childhood abuse x positive psychotic-like experiences interaction.	p=0.004; β=0.27, SE=0.10

Author	Sample size	Candidate G	Candidate E	Outcome Variable	Results	Statistics
Husted et al 2010	98 broadly defined SZ, 79 narrowly defined SZ, 86 siblings	*NOS1AP*	Childhood trauma	SZ	Narrowly defined SZ more likely to have a history of early trauma than their unaffected family members (similar results after controlling for NOS1AP).	OR=4.17; 95% CI=1.52, 11.44
Muntjeswerff et al 2011	742 SZ, 884 HC	*MTHFR 677* C>T	Winter birth	SZ	No winter period x MTHFR 677- T/T x SZ interaction.	p=0.744; OR=0.90
Chotai et al 2003	954 UPAD, BPAD, and SZ 395 HC	*TPH, 5-HTTLPR* and *DRD4*	Seasonality of birth	Season of birth variations in UPAD, BPAD and SZ	-TPH-A allele associated with one-cyclic season variation in women controls and men with BPAD. -DRD4 7-repeat associated with one-cyclic season variation in SZ women. -5-HTTPLR s allele associated with one-cyclic season variation in men with UPAD.	p=0.05 p=0.01 p=0.01
Tochigi et al 2002	110 SZ, 493 HC	*HLA-A24* and A26	Seasonality of birth	Association between HLA-A and birth-season in SZ	No association between winter birth (December-March) and A24/A26 SZ	p=0.6/0.4; $\chi^2(1)$=0.4/0.7
Narita et al 2000	60 SZ + HLA-DR1, 307 SZ no HLA-DR1	*HLA-DR1*	Seasonality of birth	Association between HLA-DR1and winter birth in SZ	HLA-DR1 associated with winter births in patients.	p=0.003; $\chi^2(1)$=8.64
Haukvik et al 2010	54 SZ, 53 HC	32 SNPs in *BDNF, DTNBP1, GRM3* and *NRG1*	Obstetric Complications (OCs)	Hippocampal volume	-GRM3 rs13242038 associated with severe OCs on hippocampal volume. -No significant interaction with SZ	$p_{diagnosis \times OCs}$=0.25 $p_{diagnosis \times OCs \times hemisphere}$=0.77
Nicodemus et al 2008	116 SZ spectrum disorders, 134 HC	*AKT1, BDNF, CAPON, CHRNA7, COMT, DTNBP1, GAD1, GRM3, NOTCH4, NRG1, PRODH, RGS4, TNF-alpha*	Obstetric Complications (OCs)	SZ	Interactions between serious OCs and: - AKT1 rs1130233. - BDNF rs2049046 and rs76882600. - DTNBP1 rs875462 - GRM3 rs7808623 - No GxE interaction in controls.	p=0.031; OR=3.97 p=0.019/0.015; OR=12.45 p=0.031; OR=9.49 p=0.061; OR=3.39
Kéri et al 2009	200 SZ	*NRG1*	Psychosocial stress	Unusual thoughts	-Two-way interaction between genotype and family interaction type. -T/T genotype associated with unusual thoughts during conflict-related interactions. -No association during neutral interactions.	p<0.0001; F(2,197)=17.98 p<0.0001 p=0.5

Author	Sample size	Candidate G	Candidate E	Outcome Variable	Results	Statistics
Simons et al 2009	579 young adult female twins (general population)	*COMT* Val(158)Met *BDNF* Val(66)Met	Stress (ESM)	Feelings of paranoia in daily life (ESM)	-COMT:Val/Val x "event stress" x paranoia interaction. No interaction with "social stress". -BDNF: Val/Met x "event stress" x paranoia interaction. No interaction with "social stress"	p=0.002; β=0.05 p=0.10; β=0.02 p<0.001; β=0.04 p=0.33; β=0.05
vanWinkel et al 2008	31 psychotic disorder + cannabis, 25 HC + cannabis	*COMT* Val(158)Met	Stress (ESM)	Psychotic experiences (ESM)	-Significant Met/Met x ESM stress x psychotic experiences interaction -Similar results for ESM delusions. -No interaction in controls.	p=<0.001; β=0.77 p=0.01; χ^2=12.4 p=0.20; χ^2=3.3

Table 3. AKT1 = Serine-threonine protein kinase; *BDNF* = Brain-derived neurotrophic factor; *CB1* = Cannabinoid receptor type 1; *CAPON* = Carboxyl-terminal PDZ ligand of neuronal nitric oxide; *CHRNA7* = Neuronal acetylcholine receptor subunit alpha-7; *COMT* = Catechol-O-methyltransferase; *DTNBP1* = dystrobrevin-binding protein 1; ESM = Experience Sampling Method. *GAD1* = Glutamate decarboxylase 1; *GRM3* = Metabotropic glutamate receptor 3; HC = Healthy Controls; *HLA* = Human Leukocyte Antigen; *MTHFR* = Methylenetetrahydrofolatereductase; *NOTCH4* = Neurogenic locus notch homolog protein 4; *NOS1AP* = Carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase protein; *NRG1* = Neuregulin; *PRODH* = Prolinedehydrogenase; *RGS4* = Regulator of G protein signaling 4; SZ = Patients with Schizophrenia; *TNF-alpha* = Tumor necrosis factor.

The particulate nature of the molecular element to GxE interaction means that genetically pre-ordained outcomes could in theory be averted. The extent to which this is true depends on the effect sizes involved and whether the proposed intervention can be made in timely fashion. This has positive implications for public health. For example in some circumstances, it may be preferable to eliminate the environmental risk component altogether, rather than attempt the more tedious task of targeting genetic risk groups for a given intervention. Phenotype expression is normally suppressed as risk-inducing environmental exposures become scarce; this correlates with a decline in heritability and impacts on the number of diagnosed cases. The contextual nature of heritability can be exploited through use of the 'exposure only' study design [87], which facilitates the detection of environmentally-sensitive genetic variation. This approach is particularly powerful at the genome-wide level. The success of such strategies is determined by the extent of GxE contribution to the heritability of a given disorder.

A good illustration of the relationship between exposure and heritability comes from a US study that compared interstate influences on the heritability of teenage nicotine use. It was found that heavier state control of tobacco availability, through a combination of higher taxation, lower advertising and controlled vending machine supply, resulted in lower levels of detectable genetic influence on daily smoking [88]. The high incidence of Schizophrenia could benefit from interventions in several areas of public policy. First and foremost would be those policies that made it more difficult to acquire cannabis, as this could reduce rates of Schizophrenia within genetically-prone sub-populations.

6. Methodological constraints in GxE research

There are lingering doubts about the experimental validity of many GxE findings reported in the literature. This is in contrast to the renewed sense of optimism about genetic association, a method which focuses on direct gene effects rather than the consequences of their interactions. Genetic association studies now have GWAS to look to as a methodological reference point [4,38,40,41,50,56,89], and it is out of the technological infrastructure supporting GWAS, that two competing theories (they may not be mutually exclusive) regarding the genetic architecture of Schizophrenia have emerged: The Common Disease-Common Variant and Multiple Rare Variant hypotheses [4,9,10,58]. The latter of these will be explored in great detail through new sequencing initiatives already underway for Schizophrenia (for example, the UK10K study: www.uk10k.org/goals.html).

A combination of meticulous study design, unprecedented sample sizes and good governance over methodological practice [90], now mean that replication is no longer the rarity it once was for genetic association studies (see table 3). This is in part due to the fact that genetic association studies are becoming more methodologically homogeneous; many of the rigorous methodological practices and standards routinely implemented in GWAS research (internally validated findings, population stratification, etc) have also been widely adopted by studies whose scope does not extend beyond individual candidate genes.

In contrast, the diversity of methodologies and standards used in GxE research has remained stubbornly heterogeneous to date; the multitude of study designs used to follow-up new discoveries, has seen only varying levels of success [91]. Longitudinal studies sit very high within the complex methodological hierarchy of epidemiological designs, but even they are failing to provide the swift resolutions hoped for, to ongoing research questions of high importance (eg. Caspi vs Zammit) [92,93].

Several recurring factors limit the success rate of replication attempts in GxE research. These include:

- Measurement error
- The distribution of genotypes and exposure
- The effect size
- Sample size

The next section is dedicated to exploring each of these aspects in greater detail.

6.1 Measurement error

Arguably the most replicated GxE finding in Psychiatry belongs to the field of Depression, and involves the short allele of the Serotonin transporter gene (*5HTTLPR*) and Stressful Life Events (SLE), which interactively augment the risk of Depression [94]. Reviews that have delved into the matter of how consistently the finding can be reproduced, have noted that there is an inverse relationship between sample sizes and the associated likelihood of replication. This appears to be due to the larger degree of measurement error (associated with exposure) inherent to large studies [95]. Small studies, which have fewer resources, shun large-scale recruitment, but place greater importance instead on maximising the accuracy with which environmental exposures are measured.

Studies in which the SLE represents a single specific source of adversity, tend to extend the interaction trend, (even if they do not strictly reach the criteria of a 'replication' study) [91]. This reaffirms the statistical importance of maintaining measurement error at low levels [96,97]. Opportunistic replication studies, typically performed using cohorts not primarily intended to address the original research question, tend to be more detrimental to replication efforts, as even variables with the same name can reflect either subtly, or grossly different constructs.

6.2 The distribution of genotypes and exposure

The issue of replication is further complicated by the fact that, depending on the frequency of the exposure, the same GxE construct may range from having:

i. no effect when the exposure is low,
ii. statistical interaction when the exposure is moderate, or
iii. a main effect when the exposure is high [91].

As genotypic frequency has a similar influence on interaction detection, it is only recommendable to attempt the reproduction of an interaction in samples where exposure and allelic frequencies compare with the original study. Additionally, power to detect interactions is optimal only when both minor allele frequencies and exposure rates are at the 50% level. Idealised distributions such as these are unlikely to occur under normal recruitment conditions, although they can be ensured by the use of selective sampling [98]. Deviation away from these two statistical optima may, along with other methodological deficiencies, compromise the overall power of a GxE study.

6.3 Effect sizes

Biological interactions need not give any statistical clues to their existence. This is demonstrated by the example of Phenylketonuria (PKU), (a syndrome that gives rise to neurodevelopmental and psychiatric symptomatologies). PKU results from a combination of allelic deficiency in the gene encoding the phenylalanine hydroxylase enzyme, and dietary exposure to phenylalanine. In this case, any statistical trace of this biological interaction is obfuscated by the ubiquitous nature of phenylalanine in the human diet.

A typical GxE analysis requires large samples to facilitate the detection of targeted effects. A wider debate surrounds how these interactions should be scaled. In order to determine the presence of an interaction, a product term is added to the regression model. In linear regression, the regression coefficient of the product term defines interaction as departure from additivity, whereas an interaction using logistic regression indicates a departure from multiplicativity [99]. An additive model is thought to best approximate the concept of biological interaction [100], though this view is heavily contentious. Meanwhile multiplicative effects, though more difficult to interpret, generally allude to larger effects on risk, and so are still predictively useful.

Biological validity remains a panacea for all GxE research, as the concept of a purely biological interaction is easy to understand and design interventions around (assuming the consequences of the interaction is large enough to merit this course of action). In contrast, inferring a mechanistic relationship out of a statistical effect, relies on conditions and

assumptions [101] that may not necessarily hold true for Schizophrenia [85]. A statistical interaction may still have great predictive value nonetheless.

The difference between these two definitions (of Biological versus Statistical Interaction) can be problematic, as there remains plenty of scope for conflict between the two. In some cases discrepancies between the two may be artefactual. For example, the logarithmic transformation inherent to the multiplicative model can cause *bona fide* interactions to disappear, or else induce them spuriously [86], (an important caveat of this strategy).

These issues have fuelled a debate about the more appropriate way to scale interaction effects eg.[85,86]. Some of the rhetoric surrounding this issue is seemingly prejudicial to the question of whether GxE research can make a positive contribution towards Schizophrenia's translational goals [85]. A key step to obtaining a definitive answer to this question will be the introduction of more systematic approaches to GxE discovery. The model for the type of approach needed is epitomised by GWAS [102,103]. In the future this will be further complemented by the genome sequencing projects now underway in Schizophrenia [104] (also see details of the UK MRC's cross-disorder sequencing initiative, the UK10K study; http://www.uk10k.org/).

6.4 Sample size

The tendency to overstate initial effect sizes results in a phenomenon known as 'winners' curse' [105].

Sample sizes in a replication study must accordingly be adjusted (upwards) to compensate for this associated loss of power. This is a practice embraced by the genetic association field of late, due to a combination of good governance [90] and a 'trickle down' of good research etiquette from GWAS, through to mainstream (candidate gene) genetic research.

A similar level of rigour is lacking from GxE research. This is worrying on two counts. Firstly, because from the outset, the power of an interaction analysis is typically lower than it is for a study of main effects [106], and secondly, the GxE field has tended to avoid facing such issues head on. This is typified by a reluctance among researchers to divulge vital information regarding statistical power in many instances [84]. Such *faux pas* are propagated by the willingness of reviewers to accept such work, without enforcing appropriate disclosure of this information.

7. 10 years of GxE research in psychiatry – A post-assessment review

A recent critical appraisal shines a spotlight on the immediate shortcomings of GxE research in the psychiatric field [84]. Its findings are still being digested by the psychiatric research community [107]. A main accusation again centres on underpowering, (described by its authors to have skewed a decade's worth of research). Face value interpretation of their calculations suggests that average effect sizes would have to be 10 times larger than those normally found in GWAS, for the small sample sizes used to be even remotely credible [84].

The problem of underpowering was found to have a bi-directional relationship with publication bias, (the tendency to only report trends that support a given hypothesis). The

authors' report outlines an interesting chain of events, initiated by the instinctive preference among journal editors for novel findings. This distortion of the literature is sustained by additional biases that favour the publication of corroborating evidence, at which point statistical considerations such as power and study design are less rigorously enforced [84]. Leniency in areas such as sample size and study design has long been self-evident in GxE research [91,95] but can, for the first time, be quantified; studies which have failed to replicate an existing discovery are, on average, 6 times larger than studies that did manage to replicate. This suggests that the sample-size threshold required for a negative finding to be published is 6x higher than that of a positive study [84].

One non-intuitive factor that such appraisals have failed to acknowledge is that samples characterised by a low *n* may also be those most immune from measurement error [91]. For the 5HTTLPR x SLE interaction alone, low measurement error has been qualitatively shown to be the single most important determinant of a successful replication [91,95]. Simulations of measurement error by Wong et al help to qualify this point [96]. They suggest that an increase in correlation with true values of 'E' from .4 to .7 can equate to as much as a 20-fold gain in sample size. It is apparent therefore, that any review of the field must take into account the fact that the problem of a small sample can, to an extent, be overcome by maximising the precision of environmental measures. These days purposefully-designed tools (eg. http://www.hsph.harvard.edu/faculty/peter-kraft/software/ or the ESPRESSO power calculator at http://www.p3gobservatory.org/powercalculator.htm) allow one to factor-in the variable precision of exposure measurement to estimations of power.

But in its defence, the Duncan-Keller assessment (a systematic assessment of 103 studies over a 10-year period) extends way beyond the Serotonin transporter. Therefore the critique is a formulation which applies to the field as a whole. Its take home message suggests that replication studies in Psychiatry currently only rarely achieve what they purport to, to a satisfactory standard.

This message is resounding, and also provides a convenient narrative for the poor progress made in bringing new findings to the clinic. At present it is largely explained by the shortage of high quality evidence entering the translational pipeline.

The crystallisation of lessons learned over the past 10 years [84] should be capitalised upon to make this a watershed moment for the application of GxE methodology in Psychiatry. However the type of cultural revolution needed can only be prompted by:

i. An all-encompassing redefinition of what constitutes methodological good practice in GxE research [107] (this could be achieved by developing something equivalent to the STREGA (*ST*rengthening the *RE*porting of *G*enetic *A*ssociations) principles, specifically for the GxE research.
ii. A consensus between journal editors, reviewers and researchers that these guidelines should be adhered to.

8. New horizons in GxE research

8.1 GxEWAS: The systematically tractable meets the biologically plausible

The archetypal approach to identifying potential GxE candidates avoids the statistical pitfalls of multiple testing, and is instead guided towards appropriate candidate regions

through a combination of biological theory and functional evidence [82]. Given our rudimentary understanding of the complexity encoded at the genomic level, it is perhaps not so surprising that the doctrine of 'biological plausibility' is often questioned. Additional scepticism is reserved for the notion that the molecular dissection of psychiatric phenotypes can be formularised [82]. This is a pertinent point, given that GWAS has shown us that the underlying biological basis of many complex and Mendelian traits is largely abstract in nature.

Advocates of the biological plausibility doctrine can rightly point to the robust experimental and analytical settings in which several of these discoveries have been made [93,94,108]. However detractors often cite the peculiarly low level of GWAS support for traditional Schizophrenia candidate gene favourites, (all of which are 'plausible' in one way or another), [109,110] to suggest the perils of a religious fixation on biological dogma [84].

The apparent discord between candidate-gene and GWAS findings is typical for most of Psychiatry, with very few exceptions [111] (convergent GWAS and candidate-gene findings in Schizophrenia are noted in table 2). If anything, GWAS has diverted attention towards less-obvious genomic points of interest, many of which lie within the non-coding domain.

Thus the non-coding genome has proved to be a rich source of pathogenic variation; approximately 90% of all GWAS findings (across disorders) originate from there. But for now, the jury is still out regarding the possible contribution of first-generation candidate genes to the risk, pathology and outcome of Schizophrenia. The delay in implementing GxEWAS studies of Schizophrenia means that the relevance of historical genetic candidates to the GxE paradigm remains untested in modern-day genome-wide protocols. It is still premature therefore, to exclude a possible wider role for some of these genes in the aetiological or pathological course of Schizophrenia.

GxEWAS studies are steadily becoming entrenched in the literature. A number of neuro-developmental and neurological phenotypes have already been investigated. These highlight interactions ranging from the effect of coffee-drinking on Parkinson's Disease, to the effect of adverse intrauterine environments on brain growth [112-114]. As this innovative branch of genomics is yet to take off in Schizophrenia, the current crop of GxE findings both in table 3 and in other areas of Psychiatry, are still yet to face the same acid test used to put the previous generation of association candidates on trial [109,110]. GxEWAS is currently one of many longer-term aspirations for policymakers in the Psychiatric Genetics community [115].

Several alternatives to standard Case-Control analysis methods will be at the disposal of the community by the time this occurs. Bayesian Case-control approaches already feature among them [116]. However the Case-only model is currently considered to be the most effective (in terms of power and efficiency) methodology for this branch of research [117,118]. The one proviso of the approach is that genes and exposure must be independent in the population from which cases are drawn [117,119]. This condition can be tested directly, by repeating the GxE analytical procedure on controls, and appropriately filtering out signals (that cross the designated threshold of significance) from the case-only study.

Post-genomic technological advances, namely the advent of micro-array technology, have led to huge increases in the scale at which genetic variation can be sampled from a genome by a single study. The abundance of this data can propel the formulation of *post-hoc* hypotheses based on biological plausibility. Useful resources that can help to inform the decision-making process include tools such as the UCSC and ENSEMBL genome browsers (http://genome.ucsc.edu/cgi-bin/hgGateway and http://www.ensembl.org/index.html). These contain a wealth of information highlighting the organisation, structure and function of the genome. Other specialist resources provide a dense functional annotation of regions that border GWAS hits (http://jjwanglab.org:8080/gwasdb/) [120].

One area in which Schizophrenia genetic research has been slow (compared to other fields such as Alzheimer's Disease), is its readiness to combine genetics with other flavours of system biology that can now be feasibly explored. This multi-level approach could provide insights about fundamental bio-mechanic processes that lie at the heart of gene-environment interaction.

One potential class of mediaries are known as Quantitative Trait Loci (QTLs). These are regulatory variants associated with control of gene-expression (eQTLs), protein levels (pQTLs) and gene activation status (methQTLs).

The ever-decreasing cost of implementing these system-based biological approaches continues to increase their accessibility. Meanwhile, whole-genome sequencing provides the means to increase both the resolution of regulatory variants across the genome, and the fuel for further biological hypotheses.

A key objective within the universal objectives of personalised medicine (to which the field of Psychiatric Genetics is also subscribed) is to enhance both the visibility and efficiency with which promising new evidence is vetted and then turned into new diagnostics and treatments. Crucially however, neither a purely biological, nor a purely systematic approach, (such as GxEWAS) can secure these goals alone. This is due to two main reasons:

- Exhausting the investigation of all plausible biological hypotheses using available genomic and enviromic data, is a slow, painstaking process that is difficult to fully automate. In any case we lack the fundamental insight about underlying biological mechanisms to assume we can become routinely successful at this.
- Meanwhile systematic methods such as GxEWAS may be too cursory. They must in any case, first confront the reasons why smaller candidate-based studies of GxE so regularly out-perform larger ones, lest the same mistakes of candidate GxE research simply end up being repeated, on a yet grander scale [91].

The many lines of derivative research resulting from GWAS in Schizophrenia collectively demonstrate how both systematic and biological candidate approaches can work in tandem [55,103,115,121,122]. Thus, an emphasis on *post-hoc* explorations of candidate pathways, genes and variants may be the best bet for turning a cursory screen of the genome (such as GxEWAS) into something that is potentially much more substantive. This kind of combinatorial approach, which marries systematic and hypothesis-led discovery through data-mining, may one day reveal (and explain) the true pervasiveness of GxE in Schizophrenia.

8.2 Strategies for data harmonisation and how this will help

Observations by Caspi [91], Uher and McGuffin [95], Vineis [97] and Wong [96] collectively highlight the challenge of balancing sample size and measurement error for optimal statistical benefit. It is in this respect that the Dunedin study (to which a disproportionate number of GxE discoveries belong) enjoys an unparalleled advantage over many of the cohorts that have since revisited the original *5HTTLPR* finding. The study combines the higher accuracy of exposure measurement often found in smaller studies, with a large sample size that is so often elusive.

A large number of replication studies do not share this same rare-but-optimal combination of properties [91,96,97]. It is this variability which may be incapacitating to the field as a whole.

Such problems can be addressed by applying greater epidemiological rigour to the collection, storage and power of genetic datasets. The rapid proliferation of biobanks in biomedical research is accompanied by the expectation that this will directly improve the quality of translational research, (and not just for Schizophrenia). Biobanks provide a means to satisfy the growing demand for high quality population data, thus they will be a key driver of genetic discovery in the future. They will also be an essential resource for validating discoveries made elsewhere.

Of course genetics is just one of many important biological areas that can be served by such resources. This is why the rapid proliferation of biobanks is vital, even for the many non-psychiatric traits that have, to a large degree, already profited from GWAS. This includes traits such as Age-related Macular Degeneration, Prostate Cancer, Coronary Heart disease and type 2 Diabetes [65].

The primary functions of a biobank include:

- Processing and storage of biological samples.
- Collection of phenotype and other data
- Facilitating statistical analysis.

A recurrent concern among commentators in the GxE field is the increased scope for measurement error in these heterogeneously-assembled datasets [91]. Additional problems may occur due to the fact that geneticists, epidemiologists, biologists and biostatisticians, often use different vocabularies [123]. Extrapolating these issues to the large number of biobanks in existence around the globe suggests that there is a need for overall governance to maximise data harmonisation. A large number of international bodies have been created for this purpose, many with overlapping functions. For example in Europe, PHOEBE (Promoting Harmonisation Of Epidemiological Biobanks in Europe), ENGAGE (European Network of Genomic and Genetic Epidemiology) P3G (Public Population Project in Genomics), are three independent organisations that provide a continent-wide consensus on procedures ranging from collection, storage and format of biological samples and associated data.

Perhaps this overlap is needed to counteract the organisational absence of other major institutions from this exercise. Regulatory bodies such as the European Medicines Agency (EMA) and The Food and Drug Administration (FDA) were at some point considered, but ultimately deemed too inherently conservative to oversee such a task [124]. Top-down

implementation of new and emerging international standards and protocols for data collection, sample acquisition, etc is managed by national biobanking initiatives, such as the UK Biobank. Policies may then be channelled down to a set of regional hubs such as the National Institute for Health Research's Biomedical Research Centres (UK). It is encouraging that Schizophrenia research is now beginning to derive the benefits of biobank-based research [125-127].

8.3 A note on methods for research synthesis

Such initiatives inevitably generate an abundance of data. A critical mass of high quality data is usually the trigger for the synthesisation of this evidence to begin. This typically uses meta-analysis, whose conventional format uses the null hypothesis (a construct of frequentist statistical theory) as its reference point. However the rationale for this becomes increasingly questionable as new evidence is added to an existing literature [128,129]. A Bayesian approach (ie. one that would allow the posterior probability of a hypothesis to be derived from prior knowledge, after taking into account new data), would allow any uncertainty about a hypothesis, to be acknowledged in an adaptive way.

The conspicuous absence of Bayesian methods from the science of data synthesis was only recently lamented, by key stakeholders involved in the process of evaluating new drugs for the UK's National Health Service [128]. Such messages may yet help to expedite the uptake of these methods, although there is already evidence of their adoption in clinical trial research [128]. These methods could widen the net used to gather new evidence, by allowing the incorporation of data from *in vivo* and cellular studies into the evaluation process. Thus Bayesian methodologies could provide an important means of channelling a wide range of functional evidence into synthesised data [130], as well as providing an alternative set of rules for assessing the validity of a hypothesis.

8.4 The future of clinical databases in psychiatric GxE research

It will soon be much easier to harvest the valuable clinical data derived out of even routine patient contact with clinical services, given that a switch-over to electronic medical records (EMRs) is now underway. The integrative blueprint for the new digital clinical age would allow a comprehensive (clinical, molecular and environmental risk profile) to be compiled for each patient. The front-end portal for this is as a personal record that follows the individual around as they move between different mental health institutions. Back-end access to such data is possible (for research purposes), but necessarily anonymised. The information itself can be processed in a way that allows even the interrogation of unstructured data (eg. clinical notes) to now be formularised (eg. see [131]). The huge potential of EMRs represents great scope for integrative research. It is anticipated that such resources will:

- Continue to drive understanding of molecular aetiology, by harnessing patients for *in silico* (bioinformatically-oriented) studies.
- Allow more efficient stratification of patients for interventions or clinical trials
- Improve the quality of genetic counselling, which will be based on a fuller, all-encompassing profile on which to base evaluations of risk, treatment outcomes and prognoses.

The true potential of the EMR model will become more apparent only when high-dimensional genetic and molecular profiling becomes economically feasible and clinically routine. This will make it practically possible to integrate a whole manner of clinical data into diagnostic/prognostic genetic research. But such times are almost already upon us, thus we do not have far to search, to find examples of how an integrative approach may work in practice. One such model is that of Ashley and colleagues [132], who recently reported a far-reaching genomic health assessment of a patient showing a strong familial indication of Coronary Heart Disease and Sudden Death Syndrome. A graphical account of the relative genetic liability for other disorders (Coronary Artery Disease, Obesity, Osteoarthritis and Type 2 Diabetes) depicts the genetic relationship between these disorders and several conditional environmental risk exposures, (stress, smoking, exercise and diet).

For Schizophrenia, a more precise account of the relationship with environmental risk factors could be achieved with the help of a new generation of instruments (questionnaires) and devices that will enable their measurement to be conducted with greater sensitivity than ever. Many examples of these have been devised for a large multi-centre study: The European network study of Gene-Environment Interaction (EUGEI) [5]. Of particular relevance is a work package entitled 'Functional Enviromics', which aims to take the elucidation of socio-environmental risk factors for Schizophrenia to a level of resolution not previously reached.

9. New horizons in pharmacogenomic research

9.1 Background

One consequence of GxE interaction is that any undesired outcomes can be averted through interventions targeted at the level of the individual, or the population, through changes in wider socio-economic policy. Primary avenues of social intervention for Schizophrenia would include redressing social inequalities [2], as well as challenging permissive attitudes to the use of illegal psychotogenic substances which, in tandem with other risk factors, help to sustain the high level of psychosis in the general population.

Meanwhile, molecular strategies for moderating or ameliorating the detrimental consequences of GxE interaction, fall within an area of personalised medicine known as Pharmacogenomics. This discipline is concerned with devising optimal therapeutic treatments for genetic sub-groups of patients. A competing goal is to minimise the risk of ill effects resulting from such treatments. Large inter-individual variability in both drug response and side-effects are the main foundation for this branch of research [133]. Much of this variability can be traced to genetic variation within key liver enzymes (the cytochrome P450 complex). It is the fate of all antipsychotic drugs to be channelled to this biological complex for breakdown (see table 4).

Of all the enzymes known to have a role in the metabolism of antipsychotic drugs, *CYP2D6* has been the most extensively characterised. This is not a great surprise, given that the protein product of this gene catalyses the breakdown of up to 25% of all pharmacological compounds. Current knowledge about functional variation within this gene alone is enough

to explain inter-individual differences in drug efficacy. For example, 2% of Caucasians and 25% of East Africans who express multiple functional CYP2D6 alleles, (ultra-rapid metabolisers) can be phenotypically distinguished on account of having the poorest levels of response to specific treatments [134,135]. Unfortunately however, current assessments of the clinical utility of pharmacogenetic testing in Schizophrenia, suggest that a heavy reliance on CYP2D6 genotyping is currently not the most beneficial way to formulate pescribing guidelines regarding the use of antipsychotic drugs [134]. A similar study of CYP2D6 (looking at Selective Serotonin Re-uptake Inhibitor treatments in Depression), recently came to a similar conclusion [136].

Table 4. Commonly used antipsychotics metabolised by CYP enzymes

Enzyme	Typical Antipsychotics	Atypical Antipsychotics
CYP2D6	**Primary metabolism**	**Primary metabolism**
	Chlorpromazine	Risperidone
	Haloperidol	
	Perphenazine	**Secondary Metabolism**
	Thioridazine	Olanzapine
		Quetiapine
	Secondary Metabolism	
	Zuclopenthixol	
CYP1A2	**Primary metabolism**	**Primary metabolism**
	Chlorpromazine	Clozapine
	Perphenazine	Olanzapine
	Thioridazine	
	Secondary Metabolism	
	Haloperidol	
	Perphenazine	
CYP3A4	**Primary metabolism**	**Primary metabolism**
	Haloperidol	Quetiapine
		Ziprasidone
		Secondary Metabolism
		Clozapine
		Olanzapine
		Risperidone

Table 4. (Adapted from reference [134])

9.2 A generalisable translation framework for GxE discovery

Poor performance of novel findings across different formulations of synthesised data represents an obvious barrier to clinical translation. But even if this obstacle can be overcome, a further series of hurdles may replace it. A clear framework now exists to prompt and signpost the long path between discovery and clinical application [137]. Implementation of the framework is marshalled by the Human Genome Epidemiology Network (HUGENET), a global collaboration of individuals and organisations whose remit is to assess the impact of genomic variation on population health. According to HUGENET, the pathway to clinical translation can be divided into four key stages (see table 5).

Table 5. The 4 phases of clinical translation

Translation Research Phase	Example	Study Approach to overcoming phase
Phase 1: Discovery and Clinical validity	eg. Reliable series of associations between a SNP and drug response	Phases I and II clinical trials; observational studies
Phase 2: Clinical Utility to Clinical guidelines	Does SNP improve drug response and what is its predictive accuracy?	Phase III clinical trials; observational studies; evidence synthesis and guidelines development
Phase 3: Implementation in Clinical practice	Explore data regarding the uptake of the SNP test in clinical settings - explore potential barriers	Dissemination and implementation research; Phase IV clinical trials
Phase 4: Public Health Impact	Does SNP improve clinical outcome in the population?	Outcomes research; Population monitoring; Phase IV clinical trials

Table 5. Table 4 shows the 4 phases of clinical translation and the critical approaches required to negotiate each one. Though initially designed to provide a translational model for pharmacogenetic research, it can also be applied in the context of GxE research. (Adapted from references [138,139])

Although this framework has been developed to support emerging new pharmacogenomic technologies, devices and treatments, its generic nature means it provides a model that is also extrapolable across genetic research (including GxE). The model adopts the ACCE (Analytical validity, Clinical Validity, Clinical Utility) and ELSI (Ethical, legal, social issues) criteria to ensure a rigorously vetted transition between phases [139]. The solid foundation provided by the framework will help to ensure that promising findings do not become 'lost in translation' [140], a problem that has characterised the last 60 years of drug development. This issue still continues to affect the industry acutely: It takes an average of 17 years for just 14% of new scientific discoveries to enter day-to-day clinical practice [137], while the cost per successful drug exceeds $1billion, after adjusting for all the failures [141].

9.3 Regulation and decision-making

Regulatory governance fulfils several objectives, the most important of which is to ensure that patients and research subjects are protected from any undesired consequences ('adverse events') of new drugs intended for the market. GxE discoveries that make it into clinical evaluation phases fall under the jurisdiction of various geographical regulatory institutions such as the European Medicines Agency (EMA) in Europe, the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK, and the Food and Drug Administration (FDA) in the US.

Adherence to the process of regulation is essential for ensuring a smooth progression through the translation scheme outlined in table 5. For instance, failing to procure accreditation for genetic tests and therapies from decision-making bodies such as the EMA and the FDA tends to adversely affect the uptake of these innovations in other global regions. This may partly explain the poor uptake of CYP2D6 and CYP2C19 genetic tests observed in a recent Danish study [142].

However, over-zealous regulation can itself create obstacles, particularly if perceived to be of no discernible benefit to patients [143]. This has potentially been the case in Europe, where the much-criticised 2001 European Union Clinical Trial Directive has caused the cost of

running clinical trials to spiral. Other knock-on effects attributed to the legislation include a 30% decline in the numbers of participants agreeing to take part in trials across Europe, over the last few years [144]. As clinical trials are an integral component within any translation scheme, such problems threaten to create a fatal bottle-neck in the pipeline, for discoveries that might otherwise have made it through the process relatively unscathed.

An overhaul of regulatory governance at national level has been proposed to circumvent this problem. In the UK, it is being done in conjunction with The National Institute for Health and Clinical Excellence (NICE), an organisation primarily responsible for assessing the cost-effectiveness, on behalf of the National Health Service (NHS), of providing new therapies and treatments. However the change of UK government means it is not even clear that there is a timetable for putting such proposals into practice [143].

As just hinted at, all novel genetic disoveries (including GxE interactions) that have safely negotiated the rigours of the validation stages shown in table 5, must still run the gauntlet of proving their overall cost-effectivenesss, before they can progress beyond validity, into utility. But new technology and treatments can only be considered to be cost-effective if their health benefits can be shown to outweigh the opportunity costs of services or treatments that they may displace [145]. When viewed in the context of the many benefits that personalised health care will bring, the additional expenses inherent to many new genomic technologies, are unlikely to present much of a barrier to widespread uptake.

10. Conclusion

Lessons of the past decade of GxE research in psychiatry (and more specifically, Schizophrenia) mean that the focus of the next should be to ensure that effort and resources already spent, or else earmarked for future investment, do not go wasted. In order to ensure this a course of greater methodological rigour should be pursued.

It would be advantageous to complement this with the encouraging array of new specialist tools, methodologies and infrastructures available, some of which are highlighted in this article. A combination of falling economic costs and increasing accessibility make this proposition the most practical and logical way forward. In the category of 'methodologies' we additionally include innovations that enable the epigenomes, transcriptomes and proteomes of Schizophrenic patients to be characterised in high-dimension. Each of these domains reflects a different dynamic (and environmentally-responsive) element within a broader biological scheme. But each also remains curiously under-represented in mainstream GxE research today. This is despite evidence to suggest they may serve a functional purpose as biomarkers of environmentally-induced pathogenesis, susceptibility, illness progression and treatment outcome [146-152]. Despite these documented examples, each discipline also faces thematic questions about how to achieve methodological best practice, given their various respective constraints [147,153,154].

Thus the current outlook would suggest that no single biological domain will have a monopoly on the clinical insights that may yet emerge out of future studies that may link genes, environment and Schizophrenia. The option to harness the various biological domains collectively, with genetics as the focal point, is promising, but currently under-resourced [155-157]. But this type of expansive approach is additionally attractive and may propel us towards fulfilling the unrealised clinical ambitions of GxE research.

11. References

[1] McGrath, J.J. & Selten, J.P. Mental health: don't overlook environment and its risk factors. Nature 454, 824 (2008).
[2] Kirkbride, J., et al. Translating the epidemiology of psychosis into public mental health: evidence, challenges and future prospects. J Public Ment Health 9, 4-14 (2010).
[3] Rujescu, D., Genius, J., Benninghoff, J. & Giegling, I. Current progress in the genetic research of schizophrenia: relevance for drug discovery? Curr Pharm Biotechnol.(2012)
[4] Ripke, S., et al. Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43, 969-976 (2011).
[5] van Os, J., Rutten, B.P. & Poulton, R. Gene-environment interactions in schizophrenia: review of epidemiological findings and future directions. Schizophr Bull 34, 1066-1082 (2008).
[6] Weis, B.K., et al. Personalized exposure assessment: promising approaches for human environmental health research. Environ Health Perspect 113, 840-848 (2005).
[7] Cardno, A.G., et al. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. Arch Gen Psychiatry 56, 162-168 (1999).
[8] Lichtenstein, P., et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 373, 234-239 (2009).
[9] Vassos, E., et al. Penetrance for copy number variants associated with schizophrenia. Hum Mol Genet 19, 3477-3481 (2010).
[10] Grozeva, D., et al. Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. Schizophr Res (2011).
[11] Dick, D.M., Riley, B. & Kendler, K.S. Nature and nurture in neuropsychiatric genetics: where do we stand? Dialogues Clin Neurosci 12, 7-23 (2010).
[12] Bergen, S.E., Gardner, C.O. & Kendler, K.S. Age-related changes in heritability of behavioral phenotypes over adolescence and young adulthood: a meta-analysis. Twin Res Hum Genet 10, 423-433 (2007).
[13] Pagan, J.L., et al. Genetic and environmental influences on stages of alcohol use across adolescence and into young adulthood. Behav Genet 36, 483-497 (2006).
[14] Agrawal, A. & Lynskey, M.T. The genetic epidemiology of cannabis use, abuse and dependence. Addiction 101, 801-812 (2006).
[15] Scherr, M., et al. Environmental risk factors and their impact on the age of onset of schizophrenia: Comparing familial to non-familial schizophrenia. Nord J Psychiatry (2011).
[16] March, D., et al. Psychosis and place. Epidemiol Rev 30, 84-100 (2008).
[17] Allardyce, J. & Boydell, J. Review: the wider social environment and schizophrenia. Schizophr Bull 32, 592-598 (2006).
[18] Cantor-Graae, E. & Selten, J.P. Schizophrenia and migration: a meta-analysis and review. Am J Psychiatry 162, 12-24 (2005).
[19] Dealberto, M.J. Ethnic origin and increased risk for schizophrenia in immigrants to countries of recent and longstanding immigration. Acta Psychiatr Scand 121, 325-339 (2010).
[20] Henquet, C., Di Forti, M., Morrison, P., Kuepper, R. & Murray, R.M. Gene-environment interplay between cannabis and psychosis. Schizophr Bull 34, 1111-1121 (2008).
[21] Arseneault, L., Cannon, M., Witton, J. & Murray, R.M. Causal association between cannabis and psychosis: examination of the evidence. Br J Psychiatry 184, 110-117 (2004).

[22] Henquet, C., Murray, R., Linszen, D. & van Os, J. The environment and schizophrenia: the role of cannabis use. Schizophr Bull 31, 608-612 (2005).
[23] Moore, T.H., et al. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. Lancet 370, 319-328 (2007).
[24] Morgan, C. & Fisher, H. Environment and schizophrenia: environmental factors in schizophrenia: childhood trauma--a critical review. Schizophr Bull 33, 3-10 (2007).
[25] Miller, B., et al. Paternal age and mortality in nonaffective psychosis. Schizophr Res 121, 218-226 (2010).
[26] Miller, B., et al. Meta-analysis of paternal age and schizophrenia risk in male versus female offspring. Schizophr Bull 37, 1039-1047 (2011).
[27] Davies, G., Welham, J., Chant, D., Torrey, E.F. & McGrath, J. A systematic review and meta-analysis of Northern Hemisphere season of birth studies in schizophrenia. Schizophr Bull 29, 587-593 (2003).
[28] Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455, 237-241 (2008).
[29] Glessner, J.T., et al. Strong synaptic transmission impact by copy number variations in schizophrenia. Proc Natl Acad Sci U S A 107, 10584-10589 (2010).
[30] Ikeda, M., et al. Copy number variation in schizophrenia in the Japanese population. Biol Psychiatry 67, 283-286 (2010).
[31] Kirov, G., et al. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum Mol Genet 18, 1497-1503 (2009).
[32] Levinson, D.F., et al. Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am J Psychiatry 168, 302-316 (2011).
[33] Need, A.C., et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet 5, e1000373 (2009).
[34] Stefansson, H., et al. Large recurrent microdeletions associated with schizophrenia. Nature 455, 232-236 (2008).
[35] Walsh, T., et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320, 539-543 (2008).
[36] Xu, B., et al. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40, 880-885 (2008).
[37] Li, J., et al. Common variants in the BCL9 gene conferring risk of schizophrenia. Arch Gen Psychiatry 68, 232-240 (2011).
[38] Shi, Y., et al. Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. Nat Genet 43, 1224-1227 (2011).
[39] Mah, S., et al. Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. Mol Psychiatry 11, 471-478 (2006).
[40] Rietschel, M., et al. Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe. Mol Psychiatry (2011).
[41] Steinberg, S., et al. Common variants at VRK2 and TCF4 conferring risk of schizophrenia. Hum Mol Genet 20, 4076-4081 (2011).
[42] Kirov, G., et al. Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. Hum Mol Genet 17, 458-465 (2008).
[43] Magri, C., et al. New copy number variations in schizophrenia. PLoS One 5, e13422 (2010).
[44] Vrijenhoek, T., et al. Recurrent CNVs disrupt three candidate genes in schizophrenia patients. Am J Hum Genet 83, 504-510 (2008).

[45] Ikeda, M., et al. Genome-wide association study of schizophrenia in a Japanese population. Biol Psychiatry 69, 472-478 (2010).
[46] O'Donovan, M.C., et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nat Genet 40, 1053-1055 (2008).
[47] Riley, B., et al. Replication of association between schizophrenia and ZNF804A in the Irish Case-Control Study of Schizophrenia sample. Mol Psychiatry 15, 29-37 (2010).
[48] Williams, H.J., et al. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. Mol Psychiatry 16, 429-441 (2011).
[49] Shi, J., et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460, 753-757 (2009).
[50] Alkelai, A., et al. Identification of new schizophrenia susceptibility loci in an ethnically homogeneous, family-based, Arab-Israeli sample. FASEB J 25, 4011-4023 (2011).
[51] Williams, H.J., et al. Most genome-wide significant susceptibility loci for schizophrenia and bipolar disorder reported to date cross-traditional diagnostic boundaries. Hum Mol Genet 20, 387-391 (2011) .
[52] Shifman, S., et al. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. PLoS Genet 4, e28 (2008).
[53] Kirov, G., et al. A genome-wide association study in 574 schizophrenia trios using DNA pooling. Mol Psychiatry 14, 796-803 (2009).
[54] Mulle, J.G., et al. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet 87, 229-236 (2010).
[55] Chen, X., et al. GWA study data mining and independent replication identify cardiomyopathy-associated 5 (CMYA5) as a risk gene for schizophrenia. Mol Psychiatry 16, 1117-1129 (2011).
[56] Yue, W.H., et al. Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. Nat Genet 43, 1228-1231 (2011) .
[57] Stefansson, H., et al. Common variants conferring risk of schizophrenia. Nature 460, 744-747 (2009).
[58] Purcell, S.M., et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752 (2009).
[59] Ingason, A., et al. A large replication study and meta-analysis in European samples provides further support for association of AHI1 markers with schizophrenia. Hum Mol Genet 19, 1379-1386 (2010)
[60] Athanasiu, L., et al. Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. J Psychiatr Res 44, 748-753 (2010)
[61] Shi, Y.Y., et al. A study of rare structural variants in schizophrenia patients and normal controls from Chinese Han population. Mol Psychiatry 13, 911-913 (2008).
[62] Vacic, V., et al. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature 471, 499-503 (2011)
[63] Green, E.K., et al. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Mol Psychiatry 15, 1016-1022 (2010).
[64] Lencz, T., et al. Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Mol Psychiatry 12, 572-580 (2007).

[65] So, H.C., Gui, A.H., Cherny, S.S. & Sham, P.C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol 35, 310-317 (2011)
[66] Eichler, E.E., et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11, 446-450 (2010)
[67] Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A (2012)
[68] Jaffee, S.R. & Price, T.S. Genotype-environment correlations: implications for determining the relationship between environmental exposures and psychiatric illness. Psychiatry 7, 496-499 (2008).
[69] Dick, D.M., et al. Gender differences in friends' influences on adolescent drinking: a genetic epidemiological study. Alcohol Clin Exp Res 31, 2012-2019 (2007).
[70] Kendler, K.S. & Karkowski-Shuman, L. Stressful life events and genetic liability to major depression: genetic control of exposure to the environment? Psychol Med 27, 539-547 (1997).
[71] Di Forti, M., et al. High-potency cannabis and the risk of psychosis. Br J Psychiatry 195, 488-491 (2009).
[72] Barkus, E.J., Stirling, J., Hopkins, R.S. & Lewis, S. Cannabis-induced psychosis-like experiences are associated with high schizotypy. Psychopathology 39, 175-178 (2006).
[73] Ferdinand, R.F., et al. Cannabis use predicts future psychotic symptoms, and vice versa. Addiction 100, 612-618 (2005).
[74] Fergusson, D.M., Horwood, L.J. & Ridder, E.M. Tests of causal linkages between cannabis use and psychotic symptoms. Addiction 100, 354-366 (2005).
[75] Veling, W., Mackenbach, J.P., van Os, J. & Hoek, H.W. Cannabis use and genetic predisposition for schizophrenia: a case-control study. Psychol Med 38, 1251-1256 (2008).
[76] McGuire, P.K., et al. Morbid risk of schizophrenia for relatives of patients with cannabis-associated psychosis. Schizophr Res 15, 277-281 (1995).
[77] Arendt, M., Mortensen, P.B., Rosenberg, R., Pedersen, C.B. & Waltoft, B.L. Familial predisposition for psychiatric disorder: comparison of subjects treated for cannabis-induced psychosis and schizophrenia. Arch Gen Psychiatry 65, 1269-1274 (2008).
[78] van Os, J., Hanssen, M., Bak, M., Bijl, R.V. & Vollebergh, W. Do urbanicity and familial liability coparticipate in causing psychosis? Am J Psychiatry 160, 477-482 (2003).
[79] Clarke, M.C., Tanskanen, A., Huttunen, M., Whittaker, J.C. & Cannon, M. Evidence for an interaction between familial liability and prenatal exposure to infection in the causation of schizophrenia. Am J Psychiatry 166, 1025-1030 (2009).
[80] Margari, F., et al. Familial liability, obstetric complications and childhood development abnormalities in early onset schizophrenia: a case control study. BMC Psychiatry 11, 60 (2011)
[81] Wicks, S., Hjern, A. & Dalman, C. Social risk or genetic liability for psychosis? A study of children born in Sweden and reared by adoptive parents. Am J Psychiatry 167, 1240-1246 (2010)
[82] Moffitt, T.E., Caspi, A. & Rutter, M. Strategy for investigating interactions between measured genes and measured environments. Arch Gen Psychiatry 62, 473-481 (2005).
[83] Tienari, P., et al. Genetic boundaries of the schizophrenia spectrum: evidence from the Finnish Adoptive Family Study of Schizophrenia. Am J Psychiatry 160, 1587-1594 (2003).

[84] Duncan, L.E. & Keller, M.C. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry 168, 1041-1049(2011)
[85] Zammit, S., Lewis, G., Dalman, C. & Allebeck, P. Examining interactions between risk factors for psychosis. Br J Psychiatry 197, 207-211(2010)
[86] Kendler, K.S. & Gardner, C.O. Interpretation of interactions: guide for the perplexed. Br J Psychiatry 197, 170-171 (2010)
[87] Kotb, M., et al. An immunogenetic and molecular basis for differences in outcomes of invasive group A streptococcal infections. Nat Med 8, 1398-1404 (2002).
[88] Boardman, J.D. State-level moderation of genetic tendencies to smoke. Am J Public Health 99, 480-486 (2009).
[89] Carrera, N., et al. Association study of nonsynonymous single nucleotide polymorphisms in schizophrenia. Biol Psychiatry 71, 169-177 (2012)
[90] Little, J., et al. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med 6, e22 (2009).
[91] Caspi, A., Hariri, A.R., Holmes, A., Uher, R. & Moffitt, T.E. Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. Am J Psychiatry 167, 509-527 (2011)
[92] Zammit, S., Owen, M.J., Evans, J., Heron, J. & Lewis, G. Cannabis, COMT and psychotic experiences. Br J Psychiatry 199, 380-385 (2011)
[93] Caspi, A., et al. Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. Biol Psychiatry 57, 1117-1127 (2005).
[94] Caspi, A., et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386-389 (2003).
[95] Uher, R. & McGuffin, P. The moderation by the serotonin transporter gene of environmental adversity in the aetiology of mental illness: review and methodological analysis. Mol Psychiatry 13, 131-146 (2008).
[96] Wong, M.Y., Day, N.E., Luan, J.A. & Wareham, N.J. Estimation of magnitude in gene-environment interactions in the presence of measurement error. Stat Med 23, 987-998 (2004).
[97] Vineis, P. A self-fulfilling prophecy: are we underestimating the role of the environment in gene-environment interaction research? Int J Epidemiol 33, 945-946 (2004).
[98] Boks, M.P., et al. Investigating gene environment interaction in complex diseases: increasing power by selective sampling for environmental exposure. Int J Epidemiol 36, 1363-1369 (2007).
[99] Knol, M.J., van der Tweel, I., Grobbee, D.E., Numans, M.E. & Geerlings, M.I. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. Int J Epidemiol 36, 1111-1118 (2007).
[100] Darroch, J. Biologic synergism and parallelism. Am J Epidemiol 145, 661-668 (1997).
[101] VanderWeele, T.J., Hernandez-Diaz, S. & Hernan, M.A. Case-only gene-environment interaction studies: when does association imply mechanistic interaction? Genet Epidemiol 34, 327-334 (2010)
[102] Engelman, C.D., et al. Detecting gene-environment interactions in genome-wide association data. Genet Epidemiol 33 Suppl 1, S68-73 (2009).
[103] Thomas, D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet 11, 259-272 (2010)

[104] Bickeboller, H., Houwing-Duistermaat, J.J., Wang, X. & Yan, X. Dealing with high dimensionality for the identification of common and rare variants as main effects and for gene-environment interaction. Genet Epidemiol 35 Suppl 1, S35-40 (2011)
[105] Kraft, P. Curses--winner's and otherwise--in genetic epidemiology. Epidemiology 19, 649-651; discussion 657-648 (2008).
[106] Kraft, P. & Hunter, D. Integrating epidemiology and genetic association: the challenge of gene-environment interaction. Philos Trans R Soc Lond B Biol Sci 360, 1609-1616 (2005).
[107] Hewitt, J.K. Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits. Behav Genet 42, 1-2 (2012)
[108] Caspi, A., et al. Role of genotype in the cycle of violence in maltreated children. Science 297, 851-854 (2002).
[109] Collins, A.L., Kim, Y., Sklar, P., O'Donovan, M.C. & Sullivan, P.F. Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results. Psychol Med 42, 607-616 (2012)
[110] Sanders, A.R., et al. No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. Am J Psychiatry 165, 497-506 (2008).
[111] Lasky-Su, J., et al. Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. Am J Med Genet B Neuropsychiatr Genet 147B, 1345-1354 (2008).
[112] Hamza, T.H., et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet 7, e1002237 (2011)
[113] Paus, T., et al. KCTD8 Gene and Brain Growth in Adverse Intrauterine Environment: A Genome-wide Association Study. Cereb Cortex (2011)
[114] Tan, A., et al. A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population. Hum Mol Genet (2012)
[115] A framework for interpreting genome-wide association studies of psychiatric disorders. Mol Psychiatry 14, 10-17 (2009).
[116] Mukherjee, B., Ahn, J., Gruber, S.B., Ghosh, M. & Chatterjee, N. Case-control studies of gene-environment interaction: Bayesian design and analysis. Biometrics 66, 934-948 (2010).
[117] Mukherjee, B., Ahn, J., Gruber, S.B. & Chatterjee, N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. Am J Epidemiol 175, 177-190 (2012) .
[118] Pierce, B.L. & Ahsan, H. Case-only genome-wide interaction study of disease risk, prognosis and treatment. Genet Epidemiol 34, 7-15 (2010).
[119] Piegorsch, W.W., Weinberg, C.R. & Taylor, J.A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13, 153-162 (1994).
[120] Li, M.J., et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res 40, D1047-1054 (2012).
[121] Chen, J., et al. Two non-synonymous markers in PTPN21, identified by genome-wide association study data-mining and replication, are associated with schizophrenia. Schizophr Res 131, 43-51(2011).
[122] Havik, B., et al. The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. Biol Psychiatry 70, 35-42 (2011).

[123] Zuvich, R.L., et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. Genet Epidemiol 35, 887-898 (2011).
[124] WellcomeTrust. Translating the potential of human population genetics research to improve the quality of health of the EU citizen. in From Biobanks to biomarkers (Wellcome Trust, 2006).
[125] Inczedy-Farkas, G., et al. [SCHIZOBANK - The Hungarian national schizophrenia biobank and its role in schizophrenia research and in personalized medicine]. Orv Hetil 151, 1403-1408 (2010).
[126] McGrath, J.J., et al. Neonatal vitamin D status and risk of schizophrenia: a population-based case-control study. Arch Gen Psychiatry 67, 889-894 (2010).
[127] Mortensen, P.B., et al. A Danish National Birth Cohort study of maternal HSV-2 antibodies as a risk factor for schizophrenia in their offspring. Schizophr Res 122, 257-263 (2010).
[128] Rawlins, M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. Lancet 372, 2152-2161 (2008).
[129] Uher, R. Forum: The case for gene-environment interactions in psychiatry. Curr Opin Psychiatry 21, 318-321 (2008).
[130] Inselman, A.L., et al. Assessment of research models for testing gene-environment interactions. Eur J Pharmacol 668 Suppl 1, S108-116.
[131] Stewart, R., et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 9, 51 (2009).
[132] Ashley, E.A., et al. Clinical assessment incorporating a personal genome. Lancet 375, 1525-1535 (2010).
[133] Wilson, J.F., et al. Population genetic structure of variable drug response. Nat Genet 29, 265-269 (2001).
[134] Fleeman, N., et al. Cytochrome P450 testing for prescribing antipsychotics in adults with schizophrenia: systematic review and meta-analyses. Pharmacogenomics J 11, 1-14 (2011) .
[135] Johansson, I., et al. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. Proc Natl Acad Sci U S A 90, 11825-11829 (1993).
[136] Thakur, M., et al. Review of evidence for genetic testing for CYP450 polymorphisms in management of patients with nonpsychotic depression with selective serotonin reuptake inhibitors. Genet Med 9, 826-835 (2007).
[137] Khoury, M.J., et al. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet Med 9, 665-674 (2007).
[138] Pirmohamed, M. Acceptance of biomarker-based tests for application in clinical practice: criteria and obstacles. Clin Pharmacol Ther 88, 862-866 (2010).
[139] Zimmern, R.L. Testing challenges: evaluation of novel diagnostics and molecular biomarkers. Clin Med 9, 68-73 (2009).
[140] Lenfant, C. Shattuck lecture--clinical research to clinical practice--lost in translation? N Engl J Med 349, 868-874 (2003).
[141] Collins, F.S. Reengineering translational science: the time is right. Sci Transl Med 3, (2011).
[142] Jurgens, G., et al. Utility and adoption of CYP2D6 and CYP2C19 genotyping and its translation into psychiatric clinical practice. Acta Psychiatr Scand (2011).

[143] Rawlins, M. A new era for UK clinical research? Lancet 377, 190-192.
[144] Perks, B. New regulations urged for UK health research. Nat Med 17, 142 (2011).
[145] Rawlins, M., Barnett, D. & Stevens, A. Pharmacoeconomics: NICE's approach to decision-making. Br J Clin Pharmacol 70, 346-349 (2010).
[146] Domenici, E., et al. Plasma protein biomarkers for depression and schizophrenia by multi analyte profiling of case-control collections. PLoS One 5, e9166 (2010).
[147] Levin, Y., et al. Global proteomic profiling reveals altered proteomic signature in schizophrenia serum. Mol Psychiatry 15, 1088-1100 (2009).
[148] Melas, P.A., et al. Epigenetic aberrations in leukocytes of patients with schizophrenia: association of global DNA methylation with antipsychotic drug treatment and disease onset. FASEB J (2012).
[149] Kurian, S.M., et al. Identification of blood biomarkers for psychosis using convergent functional genomics. Mol Psychiatry 16, 37-58 (2011).
[150] Kuzman, M.R., Medved, V., Terzic, J. & Krainc, D. Genome-wide expression analysis of peripheral blood identifies candidate biomarkers for schizophrenia. J Psychiatr Res 43, 1073-1077 (2009).
[151] Dempster, E.L., et al. Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. Hum Mol Genet 20, 4786-4796 (2011).
[152] Rutten, B.P. & Mill, J. Epigenetic mediation of environmental influences in major psychotic disorders. Schizophr Bull 35, 1045-1056 (2009).
[153] Heijmans, B.T. & Mill, J. Commentary: The seven plagues of epigenetic epidemiology. Int J Epidemiol 41, 74-78 (2012).
[154] Kumarasinghe, N., Tooney, P. & Schall, U. Finding the needle in the haystack: A review of microarray gene expression research into schizophrenia. Aust N Z J Psychiatry (2012).
[155] de Jong, S., et al. Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. Eur J Hum Genet (2012).
[156] Gibbs, J.R., et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 6, e1000952 (2010).
[157] Richards, A.L., et al. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. Mol Psychiatry 17, 193-201(2011).

Section 2

Environmental and Nutritional Issues

6

Iron Deficiency Anemia: A Public Health Problem of Global Proportions

Christopher V. Charles
University of Guelph,
Canada

1. Introduction

Iron deficiency anemia (IDA) is the most common micronutrient disorder in the world, negatively affecting the health and socio-economic wellbeing of millions of men, women, and children (Baltussen et al., 2004). According to the World Health Organization (WHO), IDA constitutes a significant public health problem requiring immediate attention from governments, researchers and healthcare practitioners (McLean et al., 2008). Iron deficiency (ID) is inherently associated with poverty, and is thus particularly prevalent in the developing world where the problem is often exacerbated by limited access to appropriate healthcare and treatment (DeMaeyer & Adiels-Tegman, 1985).

Iron deficiency and IDA result from a long term negative iron balance, culminating in decreased or exhausted iron stores (Allen, 2000; Clark 2008; Ramakrishnan & Yip, 2002). Iron, a component of every living cell, is intrinsically involved in numerous biochemical reactions in the body and is associated with oxygen transport and storage, energy production, DNA synthesis, and electron transport (Crichton et al., 2002; Theil, 2004).

Although the etiology of IDA is multifaceted, it generally results when iron demands are not met by iron absorption for any number of reasons. Individuals with IDA may have inadequate intake of iron due to poor quantity and/or quality of diet, impaired absorption or transport of iron, or chronic blood loss due to secondary disease (McLean et al., 2008).

Consequences of IDA are devastating: inhibited growth, impaired cognitive development, poor mental and motor performance, reduced work capacity, and an overall decreased quality of life (Macdougall et al., 1975; Newhouse et al., 1989; Preziosi et al., 1997; Soewondo et al., 1989; Walter et al., 1989; Zhu & Haas, 1997).

Prevention and control is typically achieved through iron fortification of food staples like flour, rice, and pasta, and/or through administration of iron supplements most often in iron pill or, more recently sprinkle form (Baltussen et al., 2004; Faqih et al., 2006; Mumtaz et al. 2000; Ramakrishnan & Yip, 2002). Although iron supplements are widely available and fortified foods constitute a major component of the diet in the developed world, access is limited in the developing world and cost if often prohibitive.

National, regional and global efforts to combat the problem of iron deficiency and IDA have garnered momentum in recent years, but the prevalence does not appear to be decreasing and the disorder remains a severe global public health problem. The current review will provide a general summary of the problem, touching upon the physiological aspects related to iron and hemoglobin, the etiology and epidemiology of IDA, and current prevention and control measures.

2. Defining iron nutritional status

Iron deficiency is defined as a condition in which there are no mobilizable iron stores, resulting from a long-term negative iron balance and leading to a compromised supply of iron to the tissues (Beutler et al., 2003). Iron status can be considered as a continuum: the ideal stage is normal iron status with varying amounts of stored iron within defined ranges; this is followed by iron deficiency, characterized by the absence of measurable iron stores; next, iron-deficient erythropoiesis shows evidence of a restricted iron supply in the absence of anemia; finally, the most significant negative consequence of ID is anemia, usually microcytic, hypochromic in nature (McLaren et al., 1983).

Anemia in general is characterized by a decrease in number of red blood cells or less than the normal quantity of hemoglobin. The condition is determined by the expected normal range of hemoglobin in a population, and is defined as existing in an individual whose hemoglobin concentration (Hb) has fallen below a threshold lying at two standard deviations below the median for a healthy population of the same demographic characteristics, including age, sex and pregnancy status (McLean et al., 2008). Anemic conditions can result from a myriad of causes that can be isolated, but more often than not co-exist. These causes include hemolysis with malaria and other infectious diseases, enzyme deficiencies, a variety of hemoglobinopathies, and other micronutrient deficiencies (McLean et al., 2008). That said, the most significant contributor to the onset of anemia worldwide is iron deficiency, and thus the terms ID, IDA, and anemia are often falsely used interchangeably. IDA represents the most severe form of iron deficiency, and has corresponding alterations in hematological laboratory values and observable signs and symptoms. Currently, the World Health Organization accepts that generally a little less than 50% of all anemias can be attributed to iron deficiency (McLean et al., 2008).

3. Biochemical and physiological importance of iron in the blood

3.1 Human iron metabolism:

Iron is important in the formation of a number of essential compounds in the body, including but not limited to hemoglobin, myoglobin, and other metalloproteins (Lynch, 1997). Most well-nourished adults in industrialized countries contain approximately 3 to 5 grams of iron, of which about 65% is in the form of hemoglobin (Bothwell, 1995). The remaining iron in the body is in the form of myoglobin, other heme compounds that promote intracellular oxidization, or is stored as ferritin in the reticuloendothelial system and cells of liver hepatocytes, bone marrow, and spleen (Frazer & Anderson 2005). Typically men have more stored iron than women, as women are often required to use

iron stores to compensate for iron loss through menstruation, pregnancy, and lactation (Bothwell, 1995).

3.2 Dietary iron sources

In food, two basic forms of iron exist: non-heme (inorganic) and heme (organic) (Bothwell, 1995; Charlton & Bothwell 1983). In an average diet, non-heme iron accounts for approximately 90% of total dietary iron content, while heme iron constitutes the remaining 10% (Bothwell et al., 1979).

Heme iron is highly bioavailable, and present in meat, fish, and poultry. In contrast, non-heme iron is not as readily bioavailable absorption is greatly influenced by diet composition (Harvey et al., 2000). Enhancers, such as ascorbic acid, and inhibitors, such as phytates and polyphenols, significantly affect inorganic iron absorption (Baynes & Bothwell, 1990, Tseng et al., 1997). Although total iron content in a meal is an important consideration, it is crucial to appreciate that the overall composition of the meal is of far greater significance for iron nutrition than the amount of total iron provided (McLean et al., 2008).

3.3 Dietary iron absorption

Dietary iron digested from food and/or supplements is absorbed by the mature villus enterocytes of the duodenum and proximal jejunum (McKie et al., 2001). Non-heme and heme iron are absorbed via different pathways, though the understanding of heme iron absorption is somewhat more limited.

Non-heme iron in ferrous form is transported across the apical membrane of enterocytes by a non-specific divalent metal transporter (DMT1) (Aisen et al., 1999, Hentze et al., 2004). Because much of the iron that enters the gastrointestinal tract is in the oxidized or ferric form, a duodenal ferric reductase (Dcytb) in the apical membrane of enterocytes reduces dietary iron prior to uptake (Latunde-Dada et al., 2002).

In contrast, heme iron molecules bind to an apical membrane protein and are absorbed intact. With the discovery of heme carrier protein 1 (HCP1), understanding has improved (Shayeghi et al., 2005). HCP1 is a polypeptide belonging to a superfamily of transporter proteins, and is predicted to have nine transmembrane domains by which heme iron is taken up. Though the mechanism is unclear, research has shown that by altering gene expression in animal models, heme absorption can be enhanced or limited by overexpressing or silencing HCP1 genes, respectively (Shayeghi et al., 2005).

Duodenal basolateral iron export into blood is mediated by the transmembrane protein ferroportin 1 (FPN1) (Zoller et al., 2001). The exact mechanism by which FPN1 functions is unclear, though it is thought to be facilitated by the ferroxidase activity of a membrane bound oxidase called hephaestin (Fleming & Bacon, 2005; Han & Kim 2007).

After moving into the plasma, iron binds to transferrin and is transported by the blood to sites of use and storage (Bailey et al., 1988). Cellular iron uptake is mediated by transferrin receptor 1 (TfR)-mediated endocytosis (Fleming & Bacon, 2005). Once inside the cell, iron has two possible fates: incorporation into iron proteins (usually as heme) or storage as ferritin for later use during times of iron deficiency (Bleackley et al., 2009).

3.4 Regulation of iron homeostasis

Since the discovery of the hormone hepcidin in 2000, the understanding of how iron homeostasis is achieved has shifted (Krause et al., 2000; Park et al. 2001). Hepcidin, a peptide hormone that is produced and predominately expressed in the liver, appears to be the master regulator of iron homeostasis in humans and other mammals (Ganz, 2003).

When iron levels are high, several regulatory molecules including hemochromatosis gene product, hemojuvelin and transferrin receptor 2, increase hepatic hepcidin expression, stimulating downstream molecular pathways. With up-regulation of hepcidin expression, iron levels are effectively regulated by binding to FPN1 which is found on the surface of iron storage cells. When iron levels are high, hepcidin causes internalization and degradation of FPN1, leading to decreased iron release from iron storage cells and a reduction in intestinal iron uptake (Dunn et al., 2007). In addition, hepcidin may also play a role in negatively regulating divalent metal transporter-1 (DMT1) and duodenal cytochrome-b (Dcytb) which are involved in intestinal iron absorption; currently, the mechanism and extent of control is unknown (Viatte et al., 2005).

In situations where iron requirements are increased, during periods of increased erythropoietic activity, anemia and hypoxia, the down-regulation of hepcidin expression is observed, though again the mechanism is not clear (Dunn et al., 2007; Pak et al., 2006; Vokurka et al., 2006)

4. Hemoglobin

4.1 Formation of hemoglobin

Hemoglobin is an allosteric protein with primary function of binding and transporting of oxygen in the blood to tissues in order to meet metabolic demands (Baldwin & Chothia, 1979). Synthesis of Hb involves a series of complex steps occurring in the erythrocytes, with production continuing through the early phases of the development and maturation of red blood cells (London et al., 1964). The coordinated production of heme, the group that mediates reversible oxygenation, and globin, which is responsible for protection of the heme group during transport, is required during synthesis (Schwartz et al., 1961).

Fully functional hemoglobin molecules consist of four globular protein subunits, each made of a protein chain that is tightly associated with a non-protein heme group (Perutz 1969, 1976; Perutz et al., 1960). The first step in the synthesis of Hb is the binding of succinyl-CoA (formed during the Krebs cycle) with glycine to form a pyrrole molecule. Next, four pyrroles combine to form protoporphyrin IX, which subsequently binds with iron to form the heme molecule. Each heme molecule then combines with a ribosomal-derived long polypeptide chain called a globin, forming a globular subunit of hemoglobin called a hemoglobin chain. Lastly, four Hb chains are loosely bound to produce a whole hemoglobin molecule. The most common form of Hb in adult humans, hemoglobin A, is a combination of two alpha and two beta chains arranged as a set of alpha-helix structural segments connected in a globin fold arrangement (Forget, 1979).

4.2 Reversible oxygenation of hemoglobin:

Aerobic metabolism is critically dependent on maintaining normal concentrations of Hb, and the protein's ability to combine with oxygen in a reversible manner is essential for normal physiological functioning (White & Beaven, 1954). Oxygen binds with Hb in the lungs during respiration and is later released in peripheral tissue capillaries in the form of molecular oxygen where the gaseous tension of the molecule is much lower than in the lungs (Campbell, 1927). This is a cooperative process as the binding affinity of hemoglobin for oxygen is increased by the oxygen saturation of the molecule (Perutz, 1980).

In addition to hemoglobin's ability to bind oxygen, the protein can also bind with carbon dioxide and carbon monoxide, though not in a cooperative manner (Christiansen et al., 1914; Hill, 1913). In the presence of carbon monoxide, hemoglobin's ability to bind with oxygen is hampered as both gases compete for the same binding site with a much greater binding affinity for carbon monoxide than oxygen (Douglas et al., 1912). As a result, small amounts of carbon monoxide can dramatically reduce the oxygen transport in the body and carbon monoxide poisoning can ensue (Hill, 1913). On the other hand, hemoglobin's ability to bind carbon dioxide is a necessary process to allow for removal of carbon dioxide and by-products from the system. Because carbon dioxide occupies a different binding site on the hemoglobin molecule, this type of ligand binding is allosteric in nature (Christiansen et al., 1914; Roughton, 1970).

4.3 Physiological control of hemoglobin levels:

The primary factor regulating the production of hemoglobin is tissue oxygenation. The peptide hormone erythropoietin (EPO), responding to a feed-back mechanism measuring blood oxygenation, is synthesized in times of decreased tissue oxygenation within 24 hours of the stimulus (Faura et al., 1969). EPO release triggers erythrocyte production in the bone marrow in an effort to achieve homeostasis of tissue oxygenation (Fandrey, 2004). As erythrocyte production increases, transferrin from plasma directly from diet and/or from iron stores enters the erythroblasts of bone marrow and is delivered to the mitochondria where heme synthesis occurs, thus inducing the formation of hemoglobin.

5. Functional consequences of iron deficiency and anemia

5.1 Cognitive development

Over the past three decades a large number of studies on the relationship between iron status and cognitive development have been conducted, often with varying results (Lozoff & Georgieff 2006). Iron and other micronutrient deficiencies often occur in the context of poverty and among individuals and families who are influenced by multiple stressors that may interfere with health and well-being, further confounding the issue. While an association between IDA and impaired cognitive development has been reported, research that takes into consideration the multi-diseased state common among individuals with IDA is needed (Lozoff & Georgieff, 2006).

Experiments employing animal models have demonstrated a key role for iron in brain development and function (Beard et al., 2006; Dallman et al., 1975; Felt & Lozoff, 1996;

Jorgenson et al., 2005; Nelson et al., 2002). Iron-containing enzymes and hemoproteins are necessary in many important development processes such as myelination, dendritogenesis, synaptogenesis, and neurotransmission. Iron deficiency disrupts these processes in a regionally specific manner depending on brain development at the time of deficiency. This disruption may lead to a variety of neurodevelopmental effects that usually do not respond to iron replenishment (Lozoff & Georgieff, 2006).

In humans, the majority of research has focused on developmental and behavioural effects of ID on infancy during 6-24 months of age. Delayed psychomotor development, cognitive performance, and social/emotional functioning have been observed in numerous studies (Grantham-McGregor & Ani, 2001; Lozoff& Georgieff, 2006).

A number of observational studies have found that children who suffered from IDA early in life continued to demonstrate lower academic performance during their school-age years. In Costa Rican children born at term and free of health problems other than moderate iron deficiency, persistence of motor differences, more grade repetition, anxiety, depression, and other social problems have been observed (Lozoff et al., 2000). When compared with children that were not anemic during infancy, these children achieved lower scores on intelligence and other cognitive performance tests upon entry into school, despite controlling for socioeconomic factors that may have acted as confounders (Lozoff et al., 1991). A recent meta-analysis estimated the long term effects on IQ to be 1.73 points lower for each 10 g/L decrease in hemoglobin during infancy (Stoltzfus et al., 2005).

The detrimental effects of iron deficiency have been ameliorated with iron supplementation. Randomized controlled trials of iron supplementation consistently show improvement in motor (Moffatt et al., 1994), social-emotional (Williams et al., 1999), and language outcomes (Stoltzfus et al. 2001).

5.2 Resistance to infection

The role that iron deficiency plays in decreased immune response has been reported in both animal and human studies (Dallman, 1987). Leukocytes (neutrophils, in particular) appear to have a reduced capacity to ingest and neutralize microorganisms (Chandra, 1973; Macdougall et al. 1975; Srikantia et al. 1976), while mitogen-stimulated lymphocytes exhibit a decreased ability to replicate (Neckers & Cossman, 1983). Additionally, depressed T-cell responses have been widely documented, with the depression proportional to the severity of iron deficiency (Chandra, 1973; Srikantia et al., 1976, Bagchi et al., 1980; Prema et al., 1982). Treatment regimens such as iron supplementation and food fortification programs have been reported to reduce morbidity from infectious disease, further implicating a role for iron in immune response (Walter et al., 1997).

5.3 Working capacity

Anemia has long been known to impair work performance, endurance, and productivity (Walker 1998). Studies in developing countries in South America (Walker, 1998; Desai et al., 1984), East Africa (Davies, 1973; Davies &Haaren, 1973), and Sri Lanka (Gardner et al., 1977) report a linear relationship between ID and work capacity. Iron supplementation studies carried out on Indonesian rubber tappers (Basta et al., 1979), and Sri Lankan (Gardner et al.,

1977) and Indonesian tea pickers (Basta et al., 1979) note significant gains in productivity following treatment of those individuals with significant IDA. One investigation conducted in China revealed that a rise of 10 g/L in Hb level was associated with an improvement in production efficiency of 14% in response to iron supplementation to treat IDA (Li et al., 1994).

A meta-analysis of 29 studies demonstrated a strong causal effect of severe and moderate IDA on aerobic work capacity in animals and humans (Haas & Brownlie, 2001). The presumed mechanism for this effect is reduced oxygen transport and reduced cellular oxidative capacity due to tissue iron deficiency (Haas & Brownlie, 2001; Davies et al., 1984). In laboratory and field trials, iron deficiency and IDA at all levels of severity also appears to affect energetic efficiency (Zhu & Haas 1997; Li et al., 1994) and endurance capacity (Edgerton et al., 1972; Rowland et al., 1988). Conversely, iron supplementation has been shown to improve endurance and aerobic work capacity in iron-depleted humans (Hinton et al., 2000; Brownlie et al., 2004; Brownlie et al., 2002).

5.4 Maternal mortality

Two meta-analyses drawing upon the same published studies reported on an association between ID and maternal mortality. In a 2001 paper, Brabin et al. suggested that there is an association between a higher risk of maternal mortality with severe anemia (Brabin et al., 2001). Stoltzfus and colleagues, using a methodologically-different analysis, corroborated these findings, suggesting that the risk of maternal mortality increased with decreasing hemoglobin levels, though not in a linear manner. Causal evidence for the role that mild or moderate anemia may play in maternal mortality is lacking (Stoltzfus et al., 2005).

In spite of these findings, a causal link between iron deficiency and mortality related to pregnancy and childbirth (ie. maternal mortality) remains unclear due to methodological concerns. To date there have been no large scale, placebo-controlled, prospective interventions to test the effect of iron supplementation on maternal mortality as large sample sizes would be required and it is considered unethical to withhold treatment from pregnant, anemic women. In addition, research in this field often does not take into consideration other possible causes of anemia and maternal mortality, such as concurrent micronutrient deficiencies, infectious disease, and other related conditions (Allen, 2000; Rush, 2000). For this reason, better observational data that controls for confounders are required (Stoltzfus et al., 2005).

5.5 Preterm delivery and growth

A negative correlation between maternal IDA with length of gestation is well established (Allen, 2001). There are currently two widely accepted biological mechanisms that explain this phenomenon (Allen, 2001). One theory suggests that anemia (leading to hypoxia) and iron deficiency (which increases serum nor-epinephrine concentrations) induces maternal and fetal stress, ultimately leading to stimulation of the production of corticotropin-releasing hormone (CRH) (Allen, 2001; Dallman, 1987; Emanuel et al., 1994). Elevated CRH is a major risk factor for preterm labour, pregnancy-induced hypertension, eclampsia, premature rupture of the membranes, maternal infection (leading to yet more

CRH synthesis), and increased fetal cortisol production (inhibiting longitudinal growth of the fetus) (Allen, 2001; Falkenberg et al., 1999; Lin et al., 1998; Linton et al., 1990, McLean et al., 1995). A second theory suggests that iron deficiency may increase oxidative damage to erythrocytes and the fetal-placental unit (Cester et al., 1994; Poranen et al., 1996).

Maternal iron deficiency with and without anemia is also strongly associated with low birth weight and impeded growth (Stoltzfus et al., 2005). While full-term infants are normally born with sufficient iron stores, infants have high iron requirements and the diets offered to infants in the developing world are frequently inadequate in terms of satisfying the iron requirements for growth. Although iron in breast milk is highly bioavailable, maternal iron reserves are depleted after 4-6 months of feeding, thus infants commonly develop iron deficiency and IDA if the diet is not altered to include a readily absorbable source of iron (Friel et al., 1990).

Iron supplementation of infants appears to ameliorate the problem of impaired growth. A number of studies conducted in Indonesia (Soewondo et al., 1989), Kenya (Latham et al. 1990), Bangladesh (Briend et al., 1990), and the United Kingdom (Aukett et al., 1986) provide evidence that iron supplementation of iron deficient children leads to improved growth.

5.6 Heavy-metal absorption

An important consequence of iron deficiency is an enhanced ability for heavy-metal uptake, leading to heavy-metal poisoning. Iron deficiency is strongly associated with an increased absorption capacity that is not specific to iron, resulting in the uptake of divalent heavy-metals like lead, cadmium, mercury and arsenic from the environment (Peraza et al., 1998). Heavy metal poisoning is a particular concern in children, as impaired cognitive development and irreversible physical and mental disability can result (Byers, 1959; Cebrian et al., 1983). For this reason, prevention of iron deficiency is important, predominantly in areas where exposure to heavy metals is common.

6. Prevalence and epidemiology

6.1 Prevalence of iron deficiency and iron deficiency anemia

Globally, nearly two billion people are affected by anemia (McLean et al., 2008). The majority of those affected live in developing countries where the problem is exacerbated by limited access to inadequate resources and appropriate treatment (Baltussen et al., 2004). IDA is unique in that it is the only nutrient deficiency which is significantly prevalent in virtually all industrialized nations as well. Currently there are no figures specifically for IDA, but it is widely accepted that approximately 50% of all cases of anemia are caused by iron deficiency (McLean et al., 2008; DeMaeyer & Adiels-Tegman, 1985). While the extent to which anemia is a problem in women and children has been widely documented, data on the prevalence of anemia in adolescents, men, and the elderly are scarce.

The level of hemoglobin concentration in the blood is used as an indicator to estimate the prevalence of anemia. Hemoglobin values that indicate the threshold for anemia have been published by the WHO and are widely accepted (Table 1) (McLean et al., 2008).

Age or gender group	Hemoglobin (g/L)
6 – 59 months	110
5 – 11 years	115
12 – 14 years	120
Non-pregnant women (>15 years)	120
Pregnant women	110
Males (>15 years)	130

Table 1. Hemoglobin levels below which anemia is present in a population.
Source: McLean et al., 2008.

Worldwide, the prevalence of anemia is highest in non-industrialized nations where prevalence is three to four times higher than developed countries (Table 2). Africa, Eastern Europe and the Western Pacific have a large burden of disease, with over 1 billion people in these three regions estimated to be anemic (McLean et al., 2008). That said, anemia in South-East Asia is more prevalent than any other region in the world, with nearly 800 million affected. While the prevalence of IDA among women and children in the developed world is lower when compared to the developing world, a high prevalence is still reported in high-risk groups, including preschool-aged children and pregnant women.

	Preschool-aged Children		Pregnant Women		Non-pregnant Women	
WHO Region	Prevalence (95% CI)	Number affected (millions)	Prevalence (95% CI)	Number affected (millions)	Prevalence (95% CI)	Number affected (millions)
Africa	67.6 (64.3-71.0)	83.5 (79.4-87.6)	57.1 (52.8-61.3)	17.2 (15.9-18.5)	47.5 (43.4-51.6)	69.9 (63.9-75.9)
Americas	29.3 (26.8-31.9)	23.1 (21.1-25.1)	24.1 (17.3-30.8)	3.9 (2.8-5.0)	17.8 (12.9-22.7)	39.0 (28.3-49.7)
South-East Asia	65.5 (61.0-70.0)	115.3 (107.3-123.2)	48.2 (43.9-52.5)	18.1 (16.4-19.7)	45.7 (41.9-49.4)	182.0 (166.9-197.1)
Europe	21.7 (15.4-28.0)	11.1 (7.9-14.4)	25.1 (18.6-31.6)	2.6 (2.0-3.3)	19.0 (14.7-23.3)	40.8 (31.5-50.1)
Eastern Mediterra-nean	46.7 (42.2-51.2)	0.8 (0.4-1.1)	44.2 (38.2-50.3)	7.1 (6.1-8.0)	32.4 (29.2-35.6)	39.8 (35.8-43.8)
Western Pacific	23.1 (21.9-24.4)	27.4 (25.9-28.9)	30.7 (28.8-32.7)	7.6 (7.1-8.1)	21.5 (20.8-22.2)	97.0 (94.0-100.0)
Global	47.4 (45.7-49.1)	293.1 (282.8-303.5)	41.8 (39.9-43.8)	56.4 (53.8-59.1)	30.2 (28.7-31.6)	468.4 (446.2-490.6)

Table 2. Anemia prevalence and number of individuals affected in pre-school aged children, pregnant women, and non-pregnant women by WHO region.
Source: (McLean et al., 2008).

Women and children are hardest hit by this nutritional disorder due to increased iron requirements during periods of growth as well as during menstruation and pregnancy. Nearly 40% of preschool children and women (aged 15-59 years) and more than 50% of all pregnant women in developing countries estimated to be anemic (McLean et al., 2008).

6.2 Etiology of iron deficiency and iron deficiency anemia

The prevalence of ID and IDA varies greatly from population to population according to a variety of host and environmental factors. The etiology of anemia is multifaceted and often several factors are at play in an anemic individual. Nutritional anemia as a result of iron deficiency is the most common cause of anemia worldwide, with approximately 50% of all cases attributed to a lack of iron in the diet. A number of host and environmental factors are associated with iron deficiency, and in more severe forms contribute to IDA as well. These include:

1. *Inadequate dietary iron intake*: Diets low in iron or diets low in adequate amounts of bioavailable iron are a major cause of IDA, particularly in non-industrialized countries. Typically, high levels of IDA are also observed in old age when dietary quality and quantity deteriorates (Clark, 2008; Fiatarone-Singh et al., 2000).
2. *Menstruation and pregnancy*: Blood losses associated with menstruation and pregnancy are common causes of ID and IDA. Typically non-menstruating women lose about 1 mg of iron per day, while menstruating women lose an additional 10 mg of iron per day during menses. Pregnancy is associated with an iron loss of approximately 1000 mg in a 55kg woman (Bothwell, 1995).
3. *Infectious disease:* In the developing world common infections which may be both chronic and recurrent are associated with blood loss leading to iron deficiency, and ultimately to IDA. Hemolytic malaria and parasitic infections such as hookworm, trichuriasis, amoebiasis, and schistosomiasis are particularly common diseases that contribute to the depletion of iron stores and often result in IDA (Oppenheimer, 2001).
4. *Interactions with medication*: Several pharmacological agents can interfere with iron uptake and/or transport leading to iron loss or defective absorption. These include H2 blockers, proton pump inhibitors, aspirin or non-steroidal anti-inflammatory drug use (Rockey & Cello, 1993).
5. *Gastrointestinal conditions*: Both acute and chronic gastrointestinal illness is associated with IDA and is an important consideration in clinical diagnosis of the condition. Duodenal or gastric ulcers, carcinoma, polyps, irritable bowel disease, erosive gastritis, celiac disease, altered hepatic function for any number of reasons, and/or compromised protein status may lead to IDA (Clark, 2008).
6. *Periods of growth*: Iron deficiency and IDA are particularly prevalent during peak periods of growth. Though full-term infants are normally born with adequate iron stores, if complementary foods containing iron are not introduced to the diet after six months of age then an infant is at risk of developing ID, and ultimately IDA. Iron requirements on a body weight basis are proportional to growth velocity, thus iron deficiency and IDA are common in preschool years and during puberty (McLean et al., 2008; Tolentino & Friedman 2007; Turner et al., 2003).
7. *Socioeconomic status*: Iron deficiency and IDA are most common among groups of low socioeconomic status for a number of reasons, including but not limited to:

malnutrition, poor education regarding health and hygiene, and greater presence of concomitant disease when compared to populations of higher socioeconomic status (Bhargava et al., 2001; Thankachan et al., 2007).

7. Prevention and control

Interventions to control iron deficiency and associated anemia are available, affordable, and sustainable (McLean et al., 2008, ACC/SCN, 2000). Food-based approaches are the most desirable and sustainable method of preventing IDA, with dietary improvement representing the most cost-effective and sustainable option. Advancements in the fortification of food staples and compliment have also shown promise. In addition, supplementation using multivitamins and vitamin complexes containing high levels of iron are accessible, though often at a higher cost than other preventive and treatment methods.

7.1 Dietary improvement

Efforts towards promoting the availability of, and access to iron-rich foods are a key prevention technique. Foods containing high levels of iron include: meat and organs from cattle, fish, and poultry, as well as non-animal foods such as legumes and green leafy vegetables (WHO, 2001).

As overall meal composition is as important as total iron content of a meal, it is important to promote the consumption of foods that enhances iron absorption and while limiting the consumption of foods that act as inhibitors (Layrisse et al. 1969; Layrisse & Martinez-Torres, 1968; Hallberg et al., 1986; Hallberg et al. 1991; Hallberg et al., 1989). Foods that enhance absorption of iron typically contain high levels of vitamins A, vitamin C, and folic acid; this includes various fruits, vegetables, and tubers. Conversely, phytates, found in cereal grains, tannins and other polyphenols found primarily in tea and coffee, and calcium from milk and milk products should be avoided where possible to limit the inhibition of iron absorption.

Typically, diets of individuals in the developing world do not provide adequate iron. In a typical South-East Asian diet consisting of rice, vegetables and spices, iron absorption was reported to be inadequate (Hallberg et al. 1974; Hallberg et al. 1977). Even with the addition of fruit, meat and fish to these simple meals, iron absorption remained lower than the estimated requirement. These findings would suggest that individuals consuming such a diet could only maintain their iron balance in a state of iron deficiency, and would therefore greatly benefit from ID and IDA treatment and prevention programs (Hallberg et al., 1974).

7.2 Iron fortification

Iron fortification involves the addition of iron to an appropriate food vehicle that is distributed widely to the general population. Fortified flour and other cereals have historically been the most commonly used (Baltussen et al., 2004; Ramakrishnan & Yip 2002). Research into self-fortification through plant breeding is also gaining momentum and in the future may have a great impact on improving nutritional status (Lucca et al., 2001).

7.2.1 Fortification in the developed world

Fortification has played an important role in the reduction of ID in the developed world since the latter half of the 20th century (Ramakrishnan & Yip 2002; Rees et al., 1985). The addition of elemental iron powder to flour and other cereals has since been commonplace, with levels of enrichment ranging from approximately 30 to 60 µg/g (Baltussen et al., 2004). In Canada, for example, it has been mandatory to enrich all white flours, enriched pastas, enriched precooked rice and certain substitute foods since 1976 (Guggenheim, 1995).

As the demand for processed foods has increased over the past 50 years, vitamins and minerals have slowly been added to an increasing array of foods. Ready-to-eat cereals in particular play an important role in daily iron intake in the Western world. Research in the last decade suggests that approximately 40% of total iron intake in women of reproductive age from the U.S. (Ramakrishnan& Yip, 2002), and approximately 78% of total iron intake in German children aged 2 to 13 years (Guggenheim, 1995) can be attributed to ready-to-eat cereals.

The fortification of infant formulas and foods provides particularly convincing evidence for the benefits of food fortification (Walter et al., 1993). Following the introduction of fortification guidelines in the U.S. in the late 1960's a clear reduction in IDA among infants and young children was noted (Yip et al., 1987).

7.2.2 Fortification in the developing world

In developing countries a much lower consumption of food from animal sources is observed and typically the overall nutritional value of the diet is lower when compared to developed nations (Yip & Ramakrishnan, 2002). In addition, both a high relative cost and a decreased availability of fortified products like cereal flours, ready-to-eat cereals, and infant formula leads to an overall decreased use of industry-prepared food that would otherwise benefit the population (Yip & Ramakrishnan, 2002). Thus, the relative absence of fortified food products from diets in developing nations could at least partly explain the high prevalence of iron deficiency and IDA and why current fortification practices have not ameliorated the situation. Research into fortifiable products that are culturally acceptable and desirable in developing nations should be conducted.

7.3 Iron supplementation

Iron supplementation is the most common and cost-effective strategy used to control ID and IDA in the developing world and is used as both a preventive measure and a treatment option (Baltussen et al., 2004). World Health Organization guidelines suggest that iron supplementation should include administration of 60 mg of iron daily with a dose of 400 µg of folic acid for women of reproductive age, 30 mg of iron and 250 µg of folic acid for school-aged children, and approximately 2 mg/kg body weight per day for preschool-aged children (McLean et al., 2008). Weekly iron supplementation also exists, though is considered to be a less effective treatment option and requires additional research and evaluation (ACC/SCN, 2001).

The majority of supplementation studies to date have examined a variety of treatments in women of reproductive age, as infants, preschool and school-aged children (Preziosi et al.,

1997; Faqih et al., 2006, Mumtaz et al. 2000; Menendez et al. 1997; Suharno et al. 1993; Schultink et al. 1995; Berger et al. 1997; Viteri, 1997). It is becoming increasingly clear that a main target group for iron supplementation in the developing world should be all women of reproductive age, regardless of pregnancy status at the time, thereby ensuring adequate iron reserves for both the mother and fetus during pregnancy and lactation (Yip & Ramakrishnan, 2002). Of concern is the relative cost of iron supplements in developing nations, coupled with issues surrounding delivery to infants and children. Other problems with iron supplementation include: undesirable side effects (including gastrointestinal irritation, black stools, and constipation); poor adherence to treatment guidelines; awareness and motivation of the target group to take supplements, often due to inadequate health and nutrition education; quality and packaging of iron supplements; and risk of iron overload if supplementation guidelines are not followed correctly (WHO, 2001).

7.4 The use of adventitious iron sources

Research conducted in the latter half of the 20th century has reported on the use of iron pots for cooking as an innovative way to reduce IDA, with the first study conducted in 1986 (Martinez & Vannucchi, 1986). Wistar rats fed a basal diet low in iron though cooked in an iron pot demonstrated comparable hemoglobin, hematocrit, protoporphyrin, serum iron, and transferrin saturation levels to those rats fed a complete diet, thus implicating the iron pot as an adventitious source of iron.

Since this time several studies have examined this supplementation technique in humans with similar findings. Experiments conducted on Ethiopian children aged 2-5 years and pre-term infants (between months 4 and 12) from Brazil reported that cooking food in iron pots led to lower rates of anemia than children whose food was cooked in non-iron pots (Adish et al. 1999; Borigato, Martinez 1998). Significantly improved hematologic values between iron pot and non-iron pot groups were noted, including increased hemoglobin, hematocrit, mean corpuscular volume, free erythrocyte protoporphyrin, and serum ferritin. In addition, the Ethiopian study indicated moderate height and weight gains in children assigned to treatment groups (Adish et al. 1999). A more recent study conducted in Malawi verifies this research, noting a reduction in iron deficiency among children and increased hemoglobin levels in adults living under malarial endemic conditions (Geerligs et al. 2003a, 2003b).

Research into the beneficial aspects of contaminant iron and adventitious iron sources, should be conducted. This supplementation technique has the possibility of providing a low-cost and sustainable way of improving dietary iron content, and may be particularly effective in the developing world where resources are limited.

8. Conclusion

Anemia is a global public health problem with serious consequences for human and socioeconomic health and development. Despite a concerted effort to improve treatment and prevention of iron deficiency and anemia in recent years, the problem does not appear to be going away.

In 2004, the Copenhagen Consensus brought together a panel of world-renowned development economists to consider and confront the ten most pressing challenges to "global welfare" that we face today (Copenhagen Consensus, 2004). Micronutrient interventions, including iron fortification, ranked at the top of the list and offered the highest benefit: cost ratio of any development intervention. These findings were confirmed in 2008, at the most recent Consensus meeting, where iron and zinc fortification were placed within the top three global challenges (Copenhagan Consensus, 2008). This prioritization of iron and other micronutrient interventions emphasizes the need for well-designed, sustainable and effective programming efforts to combat iron deficiency anemia.

The adverse effects of anemia on mortality, morbidity and development are abundantly clear. Anemia affects how individuals participate in all areas of life, including work, school and social activities, and this limits the ability to generate income and afford iron-rich sources of food, medical treatment, and school fees. In turn, this leads to constrained social and economic development, ultimately contributing to a viscous cycle of poverty that is difficult to overcome.

The widespread prevalence of anemia, both in the developed and developing worlds, is great cause for concern. The current review highlights some of the most promising research on the etiology, prevention and control of the disorder. From this, it should be clear that although we have made strides, there is still much that we do not understand about iron deficiency and anemia, especially in relation to treatment and prevention. A renewed effort to find effective ways to combat this problem is needed, as anemia is unique and complex public health crisis that is of global proportions.

9. Acknowledgements

The author wishes to thank Dr. Alastair Summerlee, President & Vice-Chancellor of the University of Guelph, Canada and Dr. Cate Dewey, Chair of Department of Population Medicine, University of Guelph, Canada for their continued support, insight, and careful editing of the chapter. The author is supported by funding from the Canadian Institutes of Health Research and the University of Guelph, Canada.

10. References

ACC/SCN United Nations (2000). Fourth Report on the World Nutrition Situation. United Nations: Geneva. Accessed on 12th February 2012 from http://www.ifpri.org/sites/default/files/pubs/pubs/books/4thrpt/4threport.pdf

Adish, A.A., Esrey, S.A., Gyorkos, T.W., Jean-Baptiste, J. & Rojhani, A. (1999). Effect of consumption of food cooked in iron pots on iron status and growth of young children: a randomised trial. *The Lancet* 353: 712-716.

Aisen, P., Wessling-Resnick, M. & Leibold, E.A. (1999). Iron metabolism. *Current Opinion in Chemical Biology* 3: 200-206.

Allen, L.H. (2001). Biological Mechanisms That Might Underlie Iron's Effects on Fetal Growth and Preterm Birth *Journal of Nutrition* 131: 581S-589S.

Allen, L.H. (2000). Anemia and iron deficiency: effects on pregnancy outcome. *American Journal of Clinical Nutrition* 71: 1280S-1284S.

Aukett, M.A., Parks, Y.A., Scott, P.H. & Wharton, B.A. (1986). Treatment with iron increases weight gain and psychomotor development. *Archives Disease in Children* 61: 849- 857.

Bailey, S., Evans, R.W., Garratt, R.C., Gorinsky, B., Hasnain, S., Horsburgh, C., Jhoti, H., Lindley, P.F. & Mydin, A. (1988). Molecular structure of serum transferrin at 3.3-.ANG. resolution *Biochemistry* 27 5804-5812.

Baldwin, J. & Chothia, C. (1979). Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *Journal of Molecular Biology* 129: 175-220.

Baltussen, R., Knai, C. & Sharan, M. (2004). Iron Fortification and Iron Supplementation are Cost-Effective Interventions to Reduce Iron Deficiency in Four Subregions of the World. *Journal of Nutrition* 134 2678-2684.

Basta, S., Karyadi, D. & Scrimshaw, N. (1979). Iron deficiency anemia and the productivity of adult males in Indonesia. *American Journal of Clinical Nutrition* 32: 916-925.

Baynes, R.D. & Bothwell, T.H. (1990). Iron Deficiency. *Annual Review of Nutrition* 10: 133-148.

Beard, J.L., Felt, B., Schallert, T., Burhans, M., Connor, J.R. & Georgieff, M.K. (2006). Moderate iron deficiency in infancy: Biology and behavior in young rats. *Behavioural Brain Research* 170: 224-232.

Berger, J., Aguayo, V.M., Tellez, W., Lujan, C., Traissac, P. & San Miguel, J.L. (1997). Weekly iron supplementation is as effective as 5 day per week iron supplementation in Bolivian school children living at high altitude. *European Journal of Clinical Nutrition* 51: 381-386.

Beutler, E., Hoffbrand, A.V. & Cook, J.D. (2003). Iron deficiency and overload. *Hematology/the Education Program of the American Society of Hematology. pp* 40-61.

Bhargava, A., Bouis, H.E. & Scrimshaw, N.S. (2001). Dietary Intakes and Socioeconomic Factors Are Associated with the Hemoglobin Concentration of Bangladeshi Women. *Journal of Nutrition* 131: 758-764.

Bleackley, M.R., Wong, A.Y.K., Hudson, D.M., Wu, C.H. & MacGillivray, R.T.A. (2009). Blood Iron Homeostasis: Newly Discovered Proteins and Iron Imbalance. *Transfusion Medicine Reviews* 23: 103-123.

Borigato, E.V.M. & Martinez, F.E. (1998). Iron Nutritional Status Is Improved in Brazilian Preterm Infants Fed Food Cooked in Iron Pots. *Journal of Nutrition* 128: 855-859.

Bothwell, T.H. (1995). Overview and mechanisms of iron regulation. *Nutrition Reviews* 53: 237-245.

Bothwell TH, Charlton RW, Cook JD, Finch CA. (1995). *Iron metabolism in man.* Blackwell Scientific Publication: Oxford, UK.

Brabin, B.J., Hakimi, M. & Pelletier, D. (2001). An Analysis of Anemia and Pregnancy-Related Maternal Mortality. *Journal of Nutrition* 131: 604S-615S.

Briend, A., Hoque, B.A. & Aziz, K.M. (1990). Iron in tubewell water and linear growth in rural Bangladesh. *Archives of Diseases of Children* 65: 224-225.

Brownlie, T.,IV, Utermohlen, V., Hinton, P.S., Giordano, C. & Haas, J.D. (2002). Marginal iron deficiency without anemia impairs aerobic adaptation among previously untrained women. *American Journal of Clinical Nutrition* 75: 734-742.

Brownlie, T.,IV, Utermohlen, V., Hinton, P.S. & Haas, J.D. (2004). Tissue iron deficiency without anemia impairs adaptation in endurance capacity after aerobic training in previously untrained women. *American Journal of Clinical Nutrition,* 79: 437-443.

Byers, R.K. (1959). Lead poisoning: Review of the Literature and Report on 45 Cases. *Pediatrics* 23: 585-603.

Campbell, J.A. (1927). Prolonged alterations of oxygen pressure in the inspired air with special reference to tissue oxygen tension, tissue carbon dioxide tension and hæmoglobin. *Journal of Physiology* 62: 211-231.

Cebrian, M.E., Albores, A., Aguilar, M. & Blakely, E. (1983). Chronic Arsenic Poisoning in the North of Mexico. *Human and Experimental Toxicology* 2: 121-133.

Cester, N., Staffolani, R., Rabini, R.A., Magnanelli, R., Salvolini, E., Galassi, R., Mazzanti, L. & Romanini, C. (1994). Pregnancy induced hypertension: a role for peroxidation in microvillus plasma membranes. *Molecular and Cellular Biochemistry* 131: 151-155.

Chandra, R.K. (1973). Reduced bactericidal capacity of polymorphs in iron deficiency.", *Archives of the Diseases of Children* 48: 864-866.

Charlton, R.W. & Bothwell, T.H. (1983). Iron Absorption. *Annual Review of Medicine* 34: 55-68.

Christiansen, J., Douglas, C.G. & Haldane, J.S. (1914). The absorption and dissociation of carbon dioxide by human blood. *Journal of Physiology* 48: 244-271.

Clark, S.F. (2008). Iron Deficiency Anemia. *Nutrition Clinical Practicum* 23:128-141. Copenhagen Consensus Centre. Copenhagen Consensus 2004. Access 12th February 2012 from: http://www.copenhagenconsensus.com/Projects/Copenhagen%20Consensus%202004-1.aspx

Crichton, R.R., Wilmet, S., Legssyer, R. & Ward, R.J. (2002). Molecular and cellular mechanisms of iron homeostasis and toxicity in mammalian cells", *Journal of Inorganic Biochemistry* 91: 9-18.

Dallman, P.R., Siimes, M.A. & Manies, E.C. (1975). Brain iron: persistent deficiency following short-term iron deprivation in the young rat. *British Journal of Haematology* 31: 209-215.

Dallman, P. (1987). Iron deficiency and the immune response. *American Journal of Clinical Nutrition* 46: 329-334.

Davies, C.T., Chukweumeka, A.C. & Van Haaren, J.P. (1973). Iron-deficiency anaemia: it effect on maximum aerobic power and responses to exercise in African males aged 17-40 years. *Clinical Science* 44: 555-562.

Davies, C.T.M. & Haaren, J.P.M.V. (1973). Effect of treatment on physiological responses to exercise in East African industrial workers with iron deficiency anaemia. *British Journal of Indian Medicine* 30: 335-340.

Davies, K.J., Donovan, C.M., Refino, C.J., Brooks, G.A., Packer, L. & Dallman, P.R. (1984). Distinguishing effects of anemia and muscle iron deficiency on exercise bioenergetics in the rat. *American Journal Physiology Endocrinology and Metabolism* 246: E535-543.

De Benoist, B., McLean. E., Egli. I, Cogswell, M (2008). Worldwide Prevalence of Anaemia 1993-2005. World Health Organization Press, Geneva. Accessed 12th February 2012 from: http://whqlibdoc.who.int/publications/2008/9789241596657_eng.pdf

DeMaeyer, E. & Adiels-Tegman, M. (1985). The prevalence of anaemia in the world", *World health statistics Quarterly/Rapport trimestriel de statistiques sanitaires mondiales* 38: 302-316.

Desai, I., Waddell, C., Dutra, S., Dutra de Oliveira, S., Duarte, E., Robazzi, M., Cevallos Romero, L., Desai, M., Vichi, F. & Bradfield, R. (1984). Marginal malnutrition and reduced physical work capacity of migrant adolescent boys in Southern Brazil. *American Journal of Clinical Nutrition* 40: 135-145.

Douglas, C.G., Haldane, J.S. & Haldane, J.B. (1912). The laws of combination of haemoglobin with carbon monoxide and oxygen. *Journal of Physiology* 44: 275-304.

Dunn, L.L., Rahmanto, Y.S. & Richardson, D.R. (2007). Iron uptake and metabolism in the new millennium. *Trends in Cell Biology* 17: 93-100.

Edgerton, V.R., Bryant, S.L., Gillespie, C.A. & Gardner, G.W. (1972). Iron Deficiency Anemia and Physical Performance and Activity of Rats. *Journal of Nutrition* 102: 381-399.

Emanuel, R.L., Robinson, B.G., Seely, E.W., Graves, S.W., Kohane, I., Saltzman, D., Barbieri, R. & Majzoub, J.A. (1994). Corticotrophin releasing hormone levels in human plasma and amniotic fluid during gestation. *Clinical Endocrinology* 40: 257-262.

Falkenberg, E.R., Davis, R.O., DuBard, M. & Parker, C.R., Jr. (1999). Effects of maternal infections on fetal adrenal steroid production. *Endocrine Research* 25: 239-249.

Fandrey, J. (2004). Oxygen-dependent and tissue-specific regulation of erythropoietin gene expression. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology* 286: R977-988.

Faqih, A.M., Kakish, S.B. & Izzat, M. (2006). Effectiveness of intermittent iron treatment of two-to six-year-old Jordanian children with iron-deficiency anemia. *Food and Nutrition Bulletin* 27: 220-227.

Faura, J., Ramos, J., Reynafarje, C., English, E., Finne, P. & Finch, C.A. (1969). Effect of Altitude on Erythropoiesis. *Blood* 33: 668-676.

Felt, B.T. & Lozoff, B. (1996). Brain Iron and Behavior of Rats are Not Normalized by Treatment of Iron Deficiency Anemia during Early Development. *Journal of Nutrition* 126: 693-701.

Fiatarone Singh, M.A., Bernstein, M.A., Ryan, A.D., O'Neill, E.F., Clements, K.M. & Evans, W.J. (2000). The effect of oral nutritional supplements on habitual dietary quality and quantity in frail elders. *The Journal of Nutrition, Health & Aging* 4: 5-12.

Fleming, R.E. & Bacon, B.R. (2005). Orchestration of Iron Homeostasis. *The New England Journal of Medicine* 352 1741-1744.

Forget, B.G. (1979). Molecular Genetics of Human Hemoglobin Synthesis. *Annals of Internal Medicine* 91: 605-616.

Frazer, D.M. & Anderson, G.J. (2005). Iron Imports. I. Intestinal iron absorption and its regulation. *American Journal Physiology Gastrointestinal Liver Physiology* 289: G631-635.

Friel, J.K., Andrews, W.L., Matthew, J.D., Long, D.R., Cornel, A.M., Cox, M. & Skinner, C.T. (1990). Iron status of very-low-birth-weight infants during the first 15 months of infancy. *Canadian Medical Association Journal* 143 733-737.

Ganz, T. (2003). Hepcidin, a key regulator of iron metabolism and mediator of anemia of inflammation. *Blood* 102 783-788.

Gardner, G., Edgerton, V., Senewiratne, B., Barnard, R. & Ohira, Y. (1977). Physical work capacity and metabolic stress in subjects with iron deficiency anemia. *American Journal of Clinical Nutrition* 30 910-917.

Geerligs, P.D., Brabin, B.J. & Omari, A.A. (2003a). Food prepared in iron cooking pots as an intervention for reducing iron deficiency anaemia in developing countries: a systematic review. *Journal of Human Nutrition and Dietetics : the official Journal of the British Dietetic Association* 16: 275-281.

Geerligs, P.P., Brabin, B., Mkumbwa, A., Broadhead, R. & Cuevas, L.E. (2003b). The effect on haemoglobin of the use of iron cooking pots in rural Malawian households in

an area with high malaria prevalence: a randomized trial. *Tropical Medicine & International Health* 8: 310-315.

Grantham-McGregor, S. & Ani, C. (2001). A Review of Studies on the Effect of Iron Deficiency on Cognitive Development in Children. *Journal of Nutrition* 131: 649S-668S.

Guggenheim, K.Y. (1995). Chlorosis: The Rise and Disappearance of a Nutritional Disease. *Journal of Nutrition* 125: 1822-1825.

Haas, J.D. & Brownlie, T., IV (2001). Iron Deficiency and Reduced Work Capacity: A Critical Review of the Research to Determine a Causal Relationship. *Journal of Nutrition* 131: 676S-690S.

Hallberg, L., Bjorn-Rasmussen, E., Rossander, L. & Suwanik, R. (1977). Iron absorption from Southeast Asian diets. II. Role of various factors that mightexplain low absorption. *American Journal of Clinical Nutrition* 30: 539-548.

Hallberg, L., Brune, M., Erlandsson, M., Sandberg, A. & Rossander-Hulten, L. (1991). Calcium: effect of different amounts on nonheme- and heme-iron absorption in humans. *American Journal of Clinical Nutrition* 53: 112-119.

Hallberg, L., Brune, M. & Rossander, L. (1989). Iron absorption in man: ascorbic acid and dose-dependent inhibition by phytate. *American Journal of Clinical Nutrition* 49: 140-144.

Hallberg, L., Brune, M. & Rossander, L. (1986). Effect of ascorbic acid on iron absorption from different types of meals. Studies with ascorbic-acid-rich foods and synthetic ascorbic acid given in different amounts with different meals. *Human Nutrition/Applied nutrition* 40: 97-113.

Hallberg, L., Garby, L., Suwanik, R. & Bjorn-Rasmussen, E (1974). Iron absorption from Southeast Asian diets. *American Journal of Clinical Nutrition* 27: 826-836.

Han, O. & Kim, E.Y. (2007). Colocalization of ferroportin-1 with hephaestin on the basolateral membrane of human intestinal absorptive cells. *Journal of Cellular Biochemistry* 101: 1000-1010.

Harvey, P.W.J., Dexter, P.B. & Darnton Hill, I. (2000). The impact of consuming iron from non-food sources on iron status in developing countries. Publications in Healthy Nutrition 3: 375–383.

Hentze, M.W., Muckenthaler, M.U. & Andrews, N.C. (2004). Balancing Acts: Molecular Control of Mammalian Iron Metabolism. *Cell* 117: 285-297.

Hill, A.V (1913). The Combinations of Haemoglobin with Oxygen and with Carbon Monoxide. I. *The Biochemical Journal* 7: 471-480.

Hinton, P.S., Giordano, C., Brownlie, T. & Haas, J.D. (2000). Iron supplementation improves endurance after training in iron-depleted, nonanemic women. *Journal of Applied Physiology* 88: 1103-1111.

Jorgenson, L.A., Sun, M., O'Connor, M. & Georgieff, M.K. (2005). Fetal iron deficiency disrupts the maturation of synaptic function and efficacy in area CA1 of the developing rat hippocampus. *Hippocampus* 15: 1094-1102.

Krause, A., Neitz, S., Mägert, H., Schulz, A., Forssmann, W., Schulz-Knappe, P. & Adermann, K. (2000). LEAP-1, a novel highly disulfide-bonded human peptide, exhibits antimicrobial activity. *FEBS letters* 480: 147-150.

Latham, M.C., Stephenson, L.S., Kinoti, S.N., Zaman, M.S. & Kurz, K.M. (1990). Improvements in growth following iron supplementation in young Kenyan school children. *Nutrition* 6: 159-165.

Latunde-Dada, G.O., Van der Westhuizen, J., Vulpe, C.D., Anderson, G.J., Simpson, R.J. & McKie, A.T. (2002). Molecular and Functional Roles of Duodenal Cytochrome B (Dcytb) in Iron Metabolism. *Blood Cells, Molecules, and Diseases* 29: 356-360.

Layrisse, M., Cook, J.D., Martinez, C., Roche, M., Kuhn, I.N., Walker, R.B. & Finch, C.A. (1969). Food Iron Absorption: A Comparison of Vegetable and Animal Foods. *Blood* 33: 430-443.

Layrisse, M. & Martinez -Torres, C. (1968). Effect of Interaction of Various Foods on Iron Absorption. *American Journal of Clinical Nutrition* 21: 1175-1183.

Li, R., Chen, X., Yan, H., Deurenberg, P., Garby, L. & Hautvast, J. (1994). Functional consequences of iron supplementation in iron-deficient female cotton mill workers in Beijing, China. *American Journal of Clinical Nutrition* 59: 908-913.

Lin, C.C. & Santolaya-Forgas, J. (1998). Current concepts of fetal growth restriction: part I. Causes, classification, and pathophysiology. *Obstetrics and Gynecology* 92: 1044-1055.

Linton, E.A., Behan, D.P., Saphier, P.W. & Lowry, P.J. (1990). Corticotropin-releasing hormone (CRH)-binding protein: reduction in the adrenocorticotropin-releasing activity of placental but not hypothalamic CRH. *Journal of Clinical Endocrinology and Metabolism* 70: 1574-1580.

London IM, Bruns GP, Karibian D. (1964). The Regulation of Hemoglobin Synthesis and the Pathogenesis of Some Hypochromic Anemias. *Medicine* 43: 789-802.

Lozoff, B., Jimenez, E. & Wolf, A. (1991). Long-term developmental outcome of infants with iron deficiency. *New England Journal of Medicine* 325: 687-694.

Lozoff, B. & Georgieff, M.K. (2006). Iron Deficiency and Brain Development. *Seminars in Pediatric Neurology* 13: 158-165.

Lozoff, B., Jimenez, E., Hagen, J., Mollen, E. & Wolf, A.W. 2000, "Poorer Behavioral and Developmental Outcome More Than 10 Years After Treatment for Iron Deficiency in Infancy", *Pediatrics,* vol. 105, no. 4, pp. e51.

Lucca, P., Hurrell, R. & Potrykus, I. (2001). Genetic engineering approaches to improve th bioavailability and the level of iron in rice grains. *Theoretical and Applied Genetics* 102: 392-397.

Lynch, S.R. (1997). Interaction of iron with other nutrients. *Nutrition Reviews* 55: 102-110.

Macdougall, L.G., Anderson, R., McNab, G.M. & Katz, J (1975). The immune response iniron-deficient children: Impaired cellular defense mechanisms with altered humoral components. *Journal of Pediatrics* 86: 833-843.

Martinez, F.E. & Vannucchi, H. (1986). Bioavailability of iron added to the diet by cooking food in an iron pot. *Nutrition Research* 6: 421-428.

McKie, A.T., Barrow, D., Latunde-Dada, G.O., Rolfs, A., Sager, G., Mudaly, E., Mudaly, M., Richardson, C., Barlow, D., Bomford, A., Peters, T.J., Raja, K.B., Shirali, S., Hediger, M.A., Garzaneh, F., & Simpson, R.J. (2001). An Iron-Regulated Ferric Reductase Associated with the Absorption of Dietary Iron. *Science* 291: 1755-1759.

McLaren, G.D., Muir, W.A. & Kellermeyer, R.W. (1983). Iron overload disorders: natural history, pathogenesis, diagnosis, and therapy. *Critical Reviews in Clinical Laboratory Sciences* 19: 205-266.

McLean, M., Bisits, A., Davies, J., Woods, R., Lowry, P. & Smith, R. (1995). A placental clock controlling the length of human pregnancy. *Nature Medicine* 1: 460-463.

Menendez, C., Kahigwa, E., Hirt, R., Vounatsou, P., Aponte, J.J., Font, F., Acosta, C.J., Schellenberg, D.M., Galindo, C.M., Kimario, J., Urassa, H., Brabin, B., Smith, T.A., Kitua, A.Y., Tanner, M. & Alonso, P.L. (1997). Randomised placebo-controlled

trial of iron supplementation and malaria chemoprophylaxis for prevention of severe anaemia and malaria in Tanzanian infants. *The Lancet* 350: 844-850.

Moffatt, M.E., Longstaffe, S., Besant, J. & Dureski, C. (1994). Prevention of iron deficiency and psychomotor decline in high-risk infants through use of iron-fortified infant formula: a randomized clinical trial. *Journal of Pediatrics* 125: 527-534.

Mumtaz, Z., Shahab, S., Butt, N., Rab, M.A. & DeMuynck, A. (2000). Daily iron supplementation is more effective than twice weekly iron supplementation in pregnant women in Pakistan in a randomized double-blind clinical trial. *Journal of Nutrition* 130: 2697-2702.

Neckers, L.M. & Cossman, J. (1983). Transferrin receptor induction in mitogen-stimulated human T lymphocytes is required for DNA synthesis and cell division and is regulated by interleukin 2. *Proceedings of the National Academy of Sciences of the United States of America* 80: 3494-3498.

Nelson, C.A., Bloom, F.E., Cameron, J.L., Amaral, D., Dahl, R.E. & Pine, D. (2002). An integrative, multidisciplinary approach to the study of brain-behavior relations in the context of typical and atypical development. *Development and Psychopathology* 14: 499-520.

Newhouse, I.J., Clement, D.B., Taunton, J.E. & McKenzie, D.C. (1989). The effects of prelatent/latent iron deficiency on physical work capacity. *Medicine and Science in Sports and Exercise* 21: 263-268.

Oppenheimer, S.J. (2001). Iron and Its Relation to Immunity and Infectious Disease. *Journal of Nutrition* 131: 616S-635S.

Pak, M., Lopez, M.A., Gabayan, V., Ganz, T. & Rivera, S. (2006). Suppression of hepcidin during anemia requires erythropoietic activity. *Blood* 108: 3730-3735.

Park, C.H., Valore, E.V., Waring, A.J. & Ganz, T. (2001). Hepcidin, a Urinary Antimicrobial Peptide Synthesized in the Liver. *Journal of Biological Chemistry* 276: 7806-7810.

Peraza, M.A., Ayala-Fierro, F., Barber, D.S., Casarez, E. & Rael, L.T. (1998). Effects of micronutrients on metal toxicity. *Environmental Health Perspectives* 106: Suppl 1. 203-216.

Perutz, M.F. (1980). Review Lecture: Stereochemical Mechanism of Oxygen Transport by Haemoglobin", *Proceedings of the Royal Society of London. Series B* 208: 135-162.

Perutz, M.F. (1976). Haemoglobin: Structure, Function and Synthesis", *British Medical Bulletin* 32: 193-194.

Perutz, M.F. (1969). Structure and function of hemoglobin. *Harvey Lectures* 63: 213-261.

Perutz, M.F.,Rossmann, M.G., Cullis, A.F., Muirhead H., Will, G. &North, A.C.. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A resolution, obtained by X-ray analysis. *Nature* 185: 416-422.

Poranen, A.K., Ekblad, U., Uotila, P. & Ahotupa, M. (1996). Lipid peroxidation and antioxidants in normal and pre-eclamptic pregnancies. *Placenta* 17: 401-405.

Prema, K., Ramalakshmi, B.A., Madhavapeddi, R. & Babu, S. (1982). Immune status of anaemic pregnant women. *British Journal of Obstetrics and Gynaecology* 89: 222-225.

Preziosi, P., Prual, A., Galan, P., Daouda, H., Boureima, H. & Hercberg, S. (1997). Effect of iron supplementation on the iron status of pregnant women: consequences for newborns. *American Journal of Clinical Nutrition* 66: 1178-1182.

Ramakrishnan, U. & Yip, R. (2002). Experiences and challenges in industrialized countries: control of iron deficiency in industrialized countries. *Journal of Nutrition* 132:. 820S-824S.

Rees, J.M., Monsen, E.R. & Merrill, J.E. 91985). Iron Fortification of Infant Foods: A Decade of Change. *Clinical Pediatrics* 24: 707-710.

Rockey, D.C. & Cello, J.P. (1993). Evaluation of the Gastrointestinal Tract in Patients with Iron-Deficiency Anemia. *New England Journal of Medicine* 329: 1691-1695.

Roughton, F.J. (1970). Some recent work on the interactions of oxygen, carbon dioxide and haemoglobin. *The Biochemical Journal* 117: 801-812.

Rowland, T.W., Deisroth, M.B., Green, G.M. & Kelleher, J.F. (1988). The effect of iron therapy on the exercise capacity of nonanemic iron-deficient adolescent runners. *American Journal of Diseases of Children* 142: 165-169.

Rush, D. (2000). Nutrition and maternal mortality in the developing world. *American Journal of Clinical Nutrition* 72: 212S-240S.

Schultink, W., Gross, R., Gliwitzki, M., Karyadi, D. & Matulessi, P. (1995). Effect of daily vs twice weekly iron supplementation in Indonesian preschool children with low iron status. *American Journal of Clinical Nutrition* 61: 111-115.

Schwartz, H.C., Goudsmit, R., Hill, R.L., Cartwright, G.E. & Wintrobe, M.M. (1961). The biosynthesis of hemoglobin from iron, protoporphyrin and globin. *Journal of Clinical Investigation* 40: 188-195.

Shayeghi, M., Latunde-Dada, G.O., Oakhill, J.S., Laftah, A.H., Takeuchi, K., Halliday, N., Khan, Y., Warley, A., McCann, F.E., Hider, R.C., Frazer, D.M., Anderson, G.J., Vulpe, C.D., Simpson, R.J. & McKie, A.T. (2005). Identification of an intestinal heme transporter. *Cel* 122: 789-801.

Soewondo, S., Husaini, M. & Pollitt, E. (1989). Effects of iron deficiency on attention and learning processes in preschool children: Bandung, Indonesia. *American Journal of Clinical Nutrition* 50: 667-674.

Srikantia, S.G., Bhaskaram, C., Prasad, J.S. & Krishnamachari, K.A.V.R. (1976). Anaemia and immune response. *The Lancet* 307: 1307-1309.

Stoltzfus, R.J., Kvalsvig, J.D., Chwaya, H.M., Montresor, A., Albonico, M., Tielsch, J.M., Savioli, L. & Pollitt, E. (2001). Effects of iron supplementation and anthelmintic treatment on motor and language development of preschool children in Zanzibar: double blind, placebo controlled study. *British Medical Journal* 323: 1389.

Stoltzfus RM, Mullany L, Black RE. (2005). Iron deficiency anaemia. In: *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors. Volume 1.* World Health Organization, Geneva pp. 163-209.

Suharno, D., Muhilal, D., Karyadi, D., West, C.E., Hautvast, J.G.A.J. & West, C.E. (1993). Supplementation with vitamin A and iron for nutritional anaemia in pregnant women in West Java, Indonesia. *The Lancet* 342: 1325-1328.

Thankachan, P., Muthayya, S., Walczyk, T., Kurpad, A.V. & Hurrell, R.F. (2007). An analysis of the etiology of anemia and iron deficiency in young women of low socioeconomic status in Bangalore, India. *Food and Nutrition Bulletin* 28: 328-336.

Theil, E.C. (2004). Iron, ferritin and nutrition. *Annual Review of Nutrition* 24: 327-343.

Tolentino, K. & Friedman, J.F. (2007). An Update on Anemia in Less Developed Countries. *American Journal of Tropical Medicine and Hygiene* 77: 44-51.

Tseng, M., Chakraborty, H., Robinson, D.T., Mendez, M. & Kohlmeier, L. (1997). Adjustment of Iron Intake for Dietary Enhancers and Inhibitors in Population Studies: Bioavailable Iron in Rural and Urban Residing Russian Women and Children. *Journal of Nutrition* 127: 1456-1468.

Turner, R.E., Langkamp-Henken, B., Littell, R.C., Lukowski, M.J. & Suarez, M.F. (2003). Comparing nutrient intake from food to the estimated average requirements shows middle-to upper-income pregnant women lack iron and possibly magnesium. *Journal of the American Dietetic Association* 103: 461-466.

Viatte, L., Lesbordes-Brion, J., Lou, D., Bennoun, M., Nicolas, G., Kahn, A., Canonne-Hergaux, F. & Vaulont, S. (2005). Deregulation of proteins involved in iron metabolism in hepcidin-deficient mice. *Blood* 105: 4861-4864.

Viteri, F.E. (1997). Iron supplementation for the control of iron deficiency in populations at risk. *Nutrition Reviews* 55: 195-209.

Vokurka, M., Krijt, J., Sulc, K. & Necas, E. (2006). Hepcidin mRNA levels in mouse liver respond to inhibition of erythropoiesis. *Physiological Research / Academia Scientiarum Bohemoslovaca* 55: 667-674.

Walker, A.R. (1998). The remedying of iron deficiency: what priority should it have? *British Journal of Nutrition* 79: 227-235.

Walter, T., Olivares, M., Pizarro, F. & Munoz, C. (1997). Iron, anemia, and infection. *Nutrition Reviews* 55: 111-124.

Walter, T., Dallman, P.R., Pizarro, F., Vebozo, L., Pena, G., Bartholmey, S.J., Hertrampf, E., Olivares, M., Letelier, A. & Arredondo, M. (1993). Effectiveness of Iron-Fortified Infant Cereal in Prevention of Iron Deficiency Anemia. *Pediatrics* 91: 976-982.

Walter, T., De Andraca, I., Chadud, P. & Perales, C.G. (1989). Iron Deficiency Anemia: Adverse Effects on Infant Psychomotor Development. *Pediatrics* 84: 7-17.

White, J.C. & Beaven, G.H. (1954). A Review of the Varieties of Human Haemoglobin in Health and Disease. *Journal of Clinical Pathology* 7: 175-200.

Williams, J., Wolff, A., Daly, A., MacDonald, A., Aukett, A. & Booth, I.W. (1999). "Iron supplemented formula milk related to reduction in psychomotor decline in infants from inner city areas: randomised study. *British Medical Journal* 318: 693-697.

World Health Organization (2001). Iron deficiency anaemia. Assessment, prevention and control: A guide for programme managers. World Health Organization, Geneva, Switzerland. Accessed 12th February 2012 from: http://www.who.int/nutrition/publications/en/ida_assessment_prevention_control.pdf

Yip, R., Binkin, N.J., Fleshood, L. & Trowbridge, F.L. (1987). Declining prevalence of anemia among low-income children in the United States. *Journal of the American Medical Association* 258: 1619-1623.

Yip, R. & Ramakrishnan, U. (2002). Experiences and challenges in developing countries. *Journal of Nutrition* 132: 827S-830S.

Zhu, Y. & Haas, J. (1997). Iron depletion without anemia and physical performance in young women. *American Journal of Clinical Nutrition* 66: 334-341.

Zoller, H., Koch, R.O., Theurl, I., Obrist, P., Pietrangelo, A., Montosi, G., Haile, D.J., Vogel, W. & Weiss, G. (2001). Expression of the duodenal iron transporters divalent-metal transporter 1 and ferroportin 1 in iron deficiency and iron overload", *Gastroenterology* 120: 1412-1419.

7

Snakebite Envenoming: A Public Health Perspective

José María Gutiérrez
Instituto Clodomiro Picado, Facultad de Microbiología,
Universidad de Costa Rica, San José,
Costa Rica

1. Introduction

Envenomings by snakebites constitute a highly relevant public health problem on a world wide basis, particularly in tropical regions of Africa, Asia and Latin America (Gutiérrez et al., 2006; WHO, 2007a). It affects mostly agricultural workers and their children living in rural settings. Thus, its highest impact occurs in poor and politically underpowered people, thus representing a 'disease of poverty' (Harrison et al., 2009) which fulfils the characteristics of a truly neglected tropical disease. Accordingly, the World Health Organization (WHO) incorporated, in 2009, snakebite envenoming in its list of neglected tropical diseases (www.who.int/neglected_disease/diseases/en). Despite the high impact of this pathology in terms of morbidity and mortality in vast regions of the world, it has received little attention from international health agencies and foundations, research agendas, and pharmaceutical companies, even when compared with other neglected diseases which have received a well deserved growing attention over the last decade (Williams et al., 2010). Such low concern for an important disease is due in part to the lack of political voice of the groups affected by snakebites, to the weakening of public health systems in many developing countries, and to the poor documentation of the actual global impact of this problem, which makes the advocacy to confront this neglected disease a difficult task. The present chapter reviews the main features associated with snakebite envenoming and its treatment, and highlights some of the most pressing tasks that need to be undertaken to confront this public health problem.

2. Assessing the actual impact of snakebite envenoming

The actual incidence and mortality associated with snakebite envenoming is poorly known, in part due to the lack of reliable information on this disease in many regions of the world. Although health statistics, based on the reports of hospital cases to health authorities, are satisfactory in some countries (for example in Brazil, de Oliveira et al., 2009), for many countries and regions this information is largely deficitary (Gutiérrez et al., 2010b; WHO, 2007a). This is in part due to the fact that health statistics are poor in many countries, and also that many people affected by snakebites do not seek medical attention and instead rely on local traditional healers, thus remaining invisible to health authorities (Habib et al., 2001; Michael et al., 2010; Otero et al., 2000; Sharma et al., 2004). Despite these limitations, a

number of studies have generated valuable information on the real impact of snakebite envenoming. Snakebites affect mainly agricultural workers and their relatives, living in poor rural settings of Africa, Asia and Latin America (Alirol et al., 2010; Chippaux, 2010; Fan & Cardoso, 1995; Warrell, 2010). Thus, it is clearly an occupational hazard. Incidence is usually higher in men than women, and children are also affected mostly due to their involvement in agricultural duties. Most bites occur in lower limbs, although bites in hands are also frequent (Alirol et al., 2010; Warrell, 2010). Incidence varies along the year, associated with the rainy season and with the timing of agricultural activities (Chippaux, 2010). Natural disasters have been associated with increments in the number of snakebites, as shown in Bangladesh during the 2007 monsoon flood (Alirol et al., 2010). Some social and ethnic groups are affected to a higher extent by snakebites, as compared with other groups. In Latin America, for instance, indigenous groups present a high incidence of snakebites (Larrick et al., 1978; Pierini et al., 1996). In addition, these groups are generally more vulnerable owing to their limited access to health services, evidencing a pattern of inequity that has implications in terms of mortality and morbidity secondary to snakebite envenomings (Gutiérrez, 2011). Moreover, these accidents fuel a vicious circle of poverty, since they have a negative impact on the working performance of agricultural workers, thus affecting the already precarious source of income for their families. Thus, in addition of being a disease of the poor (Harrison et al., 2009), snakebites worsens the economic situation of victims and their families.

A pioneer study on mortality due to this pathology was conducted by Swaroop & Grab (1954) on the basis of hospital statistics. Chippaux (1998) estimated an annual total of 5,400,000 bites, over 2,500,000 envenomings and 125,000 deaths due to snakebites. A more recent study by Kasturiratne et al. (2008) estimated a global total of envenomings ranging from 421,000 to 1,841,000, with fatalities ranging from 20,000 to 94,000. These studies presented estimations of envenomings and fatalities by regions as well. South and Southeast Asia present the highest incidence of snakebites, followed by sub-Saharan Africa (Kasturiratne et al., 2008). Likewise, these three regions have the highest numbers of fatalities. However, these estimates were based on the extrapolation of data from some regions and countries and, therefore, have limitations. When community-based surveys have been performed, the picture that emerges is one of a much higher dimension, both in terms of incidence and mortality (Snow et al., 1994; Sharma et al., 2004; Trape et al., 2001). The incidence of snakebites in specific areas can be very high. Examples are the Benue valley of Nigeria (497 per 100,000 population per year, Pugh & Theakston, 1980) and in southeastern Nepal (1,162 per 100,000 population per year, Sharma et al., 2004). A meta-analysis of snakebites in Africa suggested that the actual incidence might be 3-5 times higher than that derived from hospital statistics (Chippaux, 2011). Two recent studies further illustrate this concept. A community-based survey performed in rural Bangladesh revealed an incidence of 623.4 cases per 100,000 population per year (Rahman et al., 2010), which is much higher than the incidence derived from hospital-based statistics. Moreover, a recent study on mortality in India, which was part of a large national representative mortality survey, indicates that there are 45,900 deaths due to snakebite envenoming per year in this country (Mohapatra et al., 2011). The issue of underreporting needs to be addressed by different approaches, such as by identifying regions where underreport is more likely to occur (Hansson et al., 2010), and by performing community-based surveys in countries of high incidence of snakebites.

2.1 Beyond mortality: The impact of sequelae from snakebite envenomings

Case fatality rate in snakebite envenomings, if not properly treated, can be very high, especially in bites inflicted by highly venomous species (Sharma et al., 2004; Warrell, 2010). In addition, a percentage of people that survive develop sequelae as a consequence of envenoming. In the case of bites by viperid snakes, and by some elapids (genus *Naja*) that induce local tissue necrosis, sequelae include tissue loss and dysfunction, which may lead to amputation (S.B. Abubakar et al., 2010a; Gutiérrez & Lomonte, 2009; Otero et al., 2002; Warrell, 2010). Despite the scarcity of statistics on the incidence of sequelae following snakebite, observations in sub-Saharan Africa indicate that up to 20% of the patients, perhaps more, develop permanent physical sequelae (Pugh et al., 1980; Snow et al., 1994). Bites in the hands by viperid species are more prone to leave permanent tissue damage than bites in the lower limbs (Dart et al., 1992). Moreover, people suffering snakebites also present psychological sequelae, as clearly revealed by a recent study in Sri Lanka (S.S. Williams et al., 2011). The combination of physical and psychological consequences of snakebites has a dramatic impact on the quality of life of both patients and dependants. These are mostly poor agricultural workers whose survival depends very much on their physical and emotional stability to confront everyday hardships. In many cases, a large group of people depend on them as the only source of income. Therefore, when snakebite envenomings are analyzed using the parameter of DALYs ('disability adjusted life years') lost, the actual impact of this disease becomes more evident. It is necessary to investigate the effects of this pathology from such broader perspective, through interdisciplinary research projects involving international partnerships and networks.

In order to have a more rigorous and realistic assessment of the actual dimension of snakebite envenoming worldwide, the following tasks should be implemented: (a) Introducing compulsory notification of these envenomings. (b) Implementing the use in death certification of the specific classifier T 63.0 snake venom listed in the International Statistical Classification of Diseases and Related Health Problems (WHO, 2007b). (c) Performing well-designed epidemiological research based on health statistics and community-based surveys. (d) Supporting the training of health staff for proper record keeping on snakebite envenoming in many countries. These and related efforts will contribute to the generation of a solid body of information which will help to raise awareness on the seriousness of this problem and, at the same time, will provide decision-makers with more accurate data for the design of interventions of various sorts (Gutiérrez et al., 2010b).

3. Snake species responsible for the highest toll of envenomings

Snakes capable of inducing serious envenoming in humans are classified in the families Colubridae (*sensu lato*), Atractaspididae, Elapidae and Viperidae. These families include more than 2,600 species, although a relatively reduced number of them, mostly belonging to the families Elapidae and Viperidae, are responsible for the vast majority of snakebite envenomings worldwide (Warrell, 2010). In Asia, the most relevant species belong to the elapid genera *Bungarus* (kraits) and *Naja* (cobras) (Figure 1A), and to various species of the viperid genera *Echis*, *Daboia*, *Trimeresurus* and *Hypnale* (Warrell, 1995a). In Africa, species of *Naja* and few viperids are important in the northern countries, whereas the saw-scale viper

(*Echis ocellatus*) (Figure 1B) inflicts a heavy toll in the sub-Saharan region, together with other viperids classified in the genera *Echis* and *Bitis*, and some cobras (*Naja* sp) (WHO, 2010b). In the Americas, species of rattlesnakes (*Crotalus*) are important in North America, whereas lance-head vipers of the genus *Bothrops*, such as *B. asper* (Figure 1C) and *B. atrox*, are responsible for most snakebites in Central and South America, in addition to a number of *Bothrops* species in South America (Fan & Cardoso, 1995; Gómez & Dart, 1995; Gutiérrez, 2010).

Fig. 1A. *Naja naja* from Sri Lanka. Photo: Mark O'Shea. Reprinted from *Journal of Proteomics* 74, 1735-1767, Williams et al., copyright 2011, with permission from Elsevier.

Fig. 1B. *Echis ocellatus* from Togo. Photo: David Williams. Reprinted from *Journal of Proteomics* 74, 1735-1767, Williams et al., copyright 2011, with permission from Elsevier.

Fig. 1C. *Bothrops asper* from Costa Rica. Photo: Mahmood Sasa. From Gutiérrez et al. (2006) *PLoS Medicine* 3: e150.

In addition, some species, albeit not causing high numbers of bites, are capable of inflicting severe envenomings, such as *Lachesis* sp (bushmaster) and *Micrurus* sp (coral snakes) in the Americas (Warrell, 2004), *Atractaspis* sp (borrowing snakes) and *Dendroaspis* sp (mambas) in Africa/Middle East (WHO, 2010b), and a variety of elapid species in Australia and Papua New Guinea (White, 2010). Envenomings by colubrid species are usually not severe although fatal cases by species of the African genera *Dispholidus* and *Thelotornis* have been described (Warrell, 1995b). The taxonomy of venomous snakes is a highly dynamic field and recent modifications have been introduced in medically-relevant snake taxa (Quijada-Mascareñas & Wüster, 2010). Toxinologists, clinicians and antivenom manufacturers should be aware of these changes in taxonomy. Detailed information on the country distribution of the most important poisonous snakes is available at the WHO website http://apps.who.int/bloodproducts/snakeantivenoms/database/

4. Snake venom biochemistry and toxicology

These groups of 'advanced' snakes have acquired, through a long and complex evolutionary history (Fry et al., 2006, 2009), the ability to synthesize a toxic secretion, i.e. venom, by an exocrine gland located in the maxillary region, together with a venom delivery system based on the presence of ducts and fangs (Meier & Stocker, 1995; Vonk et al., 2008). The molecular evolution of venom toxins has involved an accelerated Darwinian process, by which genes have been duplicated and recruited in venom glands, with a concomitant process of acquisition of toxic functions based on a trend to generate mutations in sequences coding predominantly for amino acid residues located in the surface of these proteins, as well as other molecular mechanisms such as domain loss and neofunctionalization, thus generating a wide versatility in their ability to interact with diverse tissue targets (Casewell et al., 2011; Fry et al., 2006; Kini & Chan, 1999; Ohno et al., 2003). In the last decade, the use of proteomic

tools based on mass spectrometric analysis and sequence determination has allowed a detailed knowledge on the composition of venoms from many species (Calvete et al., 2007; Calvete, 2010; Fox & Serrano, 2008). Understanding the snake venom proteomes ('venomes') provides valuable information for the search of novel toxins and for the design of the most appropriate mixtures of venoms for animal immunization for antivenom production, among other applications (Calvete, 2010; Gutiérrez et al., 2009a).

Venoms from snakes of the family Elapidae comprise a high percentage of proteins of the so-called 'three finger toxin' family, which are low molecular mass (6-9 kDa) polypeptides that exert a number of actions, such as the ability to block neuromuscular junctions at the post-synaptic level by binding with very high affinity to the nicotinic cholinergic receptor of the motor end-plate in skeletal muscle fibers (Hegde et al., 2010). Some three-finger toxins are membrane-disorganizing proteins, named 'cardiotoxins' or 'cytotoxins', which disrupt the integrity of cell membranes and are likely to play a role in the tissue damage associated with envenoming by some cobras (Dufton & Hider, 1988). The venoms of *Dendroaspis* sp (mambas) contain other types of neurotoxins, i.e. dendrotoxins and fasciculins, which interfere with neuromuscular junctions by various mechanisms (Harvey, 2001, 2010). Elapid venoms are also characterized by the high abundance of phospholipases A_2 (PLA_2s), some of which are potent neurotoxins whose mechanism of action relies in the specific binding to receptors in the presynaptic nerve terminal, followed by degradation of phospholipids at the plasma membrane of these terminals, thus affecting the normal process of neurotransmitter release (Rossetto et al., 2006). Other PLA_2s induce acute muscle damage which, in the case of some sea snakes and other elapids, results in systemic myotoxicity, i.e. rhabdomyolisis, associated with myoglobinuria, hyperkalemia and acute renal failure (Gutiérrez & Ownby, 2003). Besides the predominant three-finger toxins and PLA_2s, elapid venoms also contain other proteins in low concentrations, such as cysteine-rich secretory proteins (CRISPs), cobra venom factor and other hydrolases (serine proteinases, metalloproteinases, nucleotidases) (Correa-Neto et al., 2011; Kulkeaw et al., 2007; Petras et al., 2011). The clotting disturbances induced by some Australian elapid venoms are caused by procoagulant serine proteinases which are prothrombin activators (St Pierre et al., 2005)

Venoms of snakes of the family Viperidae present large variations in their composition, but nevertheless the components showing the highest concentrations correspond to zinc-dependent metalloproteinases, PLA_2s and serine proteinases (Calvete, 2010; Fox & Serrano, 2005). In addition, these venoms contain bradykinin-potentiating peptides (BPPs), disintegrins, C-type lectin-like proteins, L-amino acid oxidase and various other enzymes (Calvete et al., 2009). Metalloproteinases are largely responsible for degradation of the basement membrane of microvessels, with the consequent hemorrhage (Escalante et al., 2011), activation of prothrombin and factor X (Kini, 2005; Tans & Rosing, 2001), thus generating the formation of microthrombi and fibrinogen depletion, i.e. defibrinogenation (Gutiérrez et al., 2010a), and degradation of the extracellular matrix (Moura-da-Silva et al., 2009), among other effects. In turn, some viperid PLA_2s induce acute muscle damage at the site of venom injection (Gutiérrez & Ownby, 2003; Lomonte et al., 2003). Some viperid PLA_2s also exert presynaptic neurotoxicity, such as the complex 'crotoxin', abundant in the venom of South American rattlesnakes (Bon, 1997). Serine proteinases are responsible for clotting disturbances, i.e. defibrinogenation, and hypotension (Serrano & Maroun, 2005). Venoms from species of the family Atractaspididae (burrowing asps) contain sarafotoxins, which are low molecular mass components that induce vasospasm leading to cardiac toxicity (Bdolah,

2010). Finally, the venoms of snakes of the polyphyletic family Colubridae have been studied to a lesser extent, but they also contain metalloproteinases, serine proteinases, PLA_2s, CRISPs and neurotoxins (Mackessy, 2002). Snake venoms present a high variability, not only between species, but also between different populations of the same species (Alape-Girón et al., 2008; Chippaux et al., 1991; Jayanthi and Gowda, 1988). Moreover, some species present a conspicuous ontogenetic variability in the composition of their venoms, such as the Central American rattlesnake *Crotalus simus* (Calvete et al., 2010a) and the lance-head viper *Bothrops asper* (Alape-Girón et al., 2008). This high variability in venom composition has evident implications for the clinical manifestations of envenoming (Warrell, 1997) and for the preparation of antivenoms (Gutiérrez et al., 2009a).

5. Clinical manifestations of envenoming

The large variation occurring in venom composition urges caution when classifying the clinical manifestations of snakebite envenoming, since important differences have been described in the clinical features in envenomings by closely-related species or even within a single species. However, there are general trends in the clinical picture of envenoming by the various groups of poisonous snakes. Envenomings by elapid species (sea snakes, tiger snakes and taipans in Australia, cobras and kraits in Asia, cobras and mambas in Africa, and coral snakes in the Americas) are usually characterized by progressive descending neurotoxic paralysis secondary to the action of pre- or post-synaptic neurotoxins at the neuromuscular junctions (Warrell, 1996, 2010; White, 2010). The most serious consequence of this effect is respiratory paralysis, which may lead to death if not properly and timely attended. In addition, envenomings by a number of elapid species are also characterized by rhabdomyolysis, which may lead to acute renal failure (Warrell, 1996). Patients envenomed by elapids in Australia and Papua New Guinea develop coagulation disturbances which may provoke bleeding (White, 2010). On the other hand, human envenomings by some cobras in Asia and Africa are not characterized by neurotoxic manifestations, but instead by local tissue necrosis (Warrell, 1995a, 1995b).

Viperid snake venoms provoke complex and often drastic local pathological effects, i.e. hemorrhage, dermonecrosis, blistering, myonecrosis and edema, always associated with pain (Gutiérrez & Lomonte, 2009; Warrell, 2004). These local manifestations may lead to permanent sequelae, such as tissue loss and dysfunction (Dart et al., 1992; Otero et al., 2002). After systemic venom distribution, and depending on the severity of the case, viperid snakebite envenomings are characterized by coagulopathies, bleeding, renal alterations and hemodynamic manifestations which may lead to cardiovascular shock and multisystem organ failure (Gutiérrez et al., 2009b; Warrell, 2004). Intravascular hemolysis might also occur, in some cases associated with microthrombi formation (Warrell, 1996). Exceptions to this general trend are envenomings by the South American and some populations of North American rattlesnakes, as well as some viperids in the Old World, which induce neurotoxicity (Azevedo-Marques et al., 2009; Ferquel et al., 2007). Despite the existence of these general trends, clinical studies highlight the complexity of snakebite envenoming, as demonstrated by the description of 'unusual' manifestations in cases by some elapids in Asia and South America (Faiz et al., 2010; Manock et al., 2008; Trinh et al., 2010). In addition, some venoms induce unique clinical features, such as the thrombotic effect described for the Caribbean viperid species *Bothrops lanceolatus* and *B. caribbaeus* (Thomas et al., 1996), and the acute hemorrhagic infarction of the pituitary in envenoming by *Daboia russelli* (Tun-Pe et al., 1987).

The severity of snakebite envenoming depends on a number of factors, such as the volume of venom injected, the size and physiological condition of the victim, and the region of the body where venom is delivered. A percentage of snakebites are not associated with venom injection ('dry bites') and, therefore, no clinical manifestations develop (Warrell, 2004). In general, bites in the head tend to be more severe than bites in the extremities, and envenoming in children are more prone to become severe. In the case of envenoming by pit vipers, bites in the hands are more likely to generate sequelae than bites in the lower limbs (Dart et al., 1992). Thus, a proper assessment of the clinical manifestations and severity of snakebites is a key element for the correct diagnosis and clinical management of these accidents.

6. Diagnosis and treatment of snakebite envenomings

6.1 Diagnosis

Identification of the offending snake is often difficult because in many settings there are various similar species and the bitten person is usually unable to differentiate between them. Even when the snake is killed and brought to the health facility, identification is not always correct. In Australia, kits have been developed for the immunodetection of venom in the bite site or in urine, thus allowing the identification of the offending snake (White, 2010). However, this is not the case in the vast majority of regions in the rest of the world. A 'syndromic approach' has been promoted for the diagnosis of the type of envenoming in various parts of the world (Ariaratnam et al., 2009; WHO, 2010b). For instance, in Central America, there are two predominant syndromes in snakebite envenomings: one presenting local pathological effects (swelling, pain, local tissue damage), clotting disturbances and bleeding, and another characterized by descending neuromuscular paralysis. The first syndrome is associated with envenomings inflicted by viperid species, whereas the second is due to envenomings by elapid species (*Micrurus* sp). This clinically-based diagnosis allows for the selection of the correct antivenom, i.e. polyvalent antivenom or anti-coral antivenom, respectively (Gutiérrez, 2010). Such syndromic approach has been advocated in other regions of the world as well, such as in sub-Saharan Africa (WHO, 2010b) and Sri Lanka (Ariaratnam et al., 2009). In large regions of the savannahs in sub-Saharan Africa, cases presenting clotting disturbances are associated with envenomings inflicted by the saw-scale viper, *Echis ocellatus* (Warrell, 1995b). In this context, a simple laboratory test known as the '20 minute whole blood clotting test' represents a useful diagnostic tool (Warrell et al., 1974). In contrast, envenomings associated with a predominantly neurotoxic picture are caused by species of neurotoxic cobras (*Naja* sp) or mambas (*Dendroaspis* sp), and envenomings characterized by local tissue damage without coagulant disturbances are induced by species of *Bitis* or by cytotoxic cobras (WHO, 2010b).

6.2 First aid in snakebite envenoming

Snakebite cases in many regions of the world are initially attended by local healers who use a wide variety of interventions, most of which are ineffective and often exert harmful effects. Examples are the use of ligatures, incisions and suction, cryotherapy, electroshock, and the administration of synthetic or natural substances (Hardy, 2009; Warrell, 2010). Other interventions, such as application of 'black stone' or suction devices are largely ineffective for the removal of venom. In addition to their harmful effects, these actions delay the

transport of patients to health centers and, therefore, jeopardize the adequate management of these cases. First aid interventions should be focused on the immobilization of the bitten extremity and the rapid transportation to clinics or other health facilities. Communities should have strategies for rapid deployment of snakebitten people to medical treatment; an example is the use of motorcycle transportation in Nepal (Alirol et al., 2010). The interaction and communication of health staff with local healers is very important, in order to promote partnerships aimed at reducing harmful interventions and guaranteeing rapid mobilization for antivenom administration. The application of pressure-immobilization, by applying a bandage and a splint to the entire bitten limb, has been used in Australia for delaying the systemic absorption of neurotoxic venoms (Sutherland et al., 1979; White, 2010). Recently, a pharmacological intervention, based on the application of an ointment containing a nitric oxide donor, aimed at reducing the lymphatic absorption of venom, has been proposed (Saul et al., 2011), and its testing in the clinical setting is pending.

6.3 Antivenoms: The key therapy of snakebite envenoming

The parenteral administration of animal-derived antivenoms constitutes the mainstay in the therapy of snakebite envenoming (WHO, 2007a, 2010a), since the development of the first antivenoms, the *serum anti-venimeux*, during the last decade of the XIXth century (Bon, 1996). Antivenoms are preparations of immunoglobulins, or immunoglobulin fragments $F(ab')_2$ or Fab, obtained by fractionating the plasma of animals immunized with snake venoms (Gutiérrez et al., 2011a; Lalloo & Theakston, 2003; WHO, 2010a). Antivenoms can be monospecific, when animals receive the venom of a single species, or polyspecific, when venoms from two or more species are injected. The majority of manufacturers use horses for immunization, although few use sheep and donkeys (Gutiérrez et al., 2011a, 2011b). In most cases, plasma fractionation involves the digestion of proteins with pepsin or, by few producers, with papain, followed by the purification of antibody fragments by salting-out with ammonium salts or caprylic acid fractionation and, in some cases, with chromatographic procedures (dos Santos et al., 1989; Grandgeorge et al., 1996; Raw et al., 1991; WHO, 2010a). Some producers fractionate plasma with caprylic acid to obtain whole IgG preparations (Gutiérrez et al., 2005; Rojas et al., 1994). A detailed description of the methods used in animal immunization and plasma fractionation for antivenom production can be found in the *WHO Guidelines for the Production, Control and Regulation of Snake Antivenom Immunoglobulins* (WHO, 2010a). There are antivenom-manufacturing laboratories in every continent (a complete list can be found in http://apps.who.int/bloodproducts/snakeantivenoms/database/). Following manufacture, antivenoms are subjected to a quality control protocol which involves physical, chemical and biological tests aimed at ensuring the efficacy and safety of these products (Gutiérrez & León, 2009; WHO, 2010a).

The ability of antivenoms to neutralize venom toxins is based on the capacity of antivenom antibodies, or antibody fragments, to bind and neutralize the most relevant toxins in a venom. It has been proposed that such neutralization is based on, at least, four mechanisms: (a) Binding of antibody paratopes to epitopes located at the pharmacologically-relevant molecular region, i.e. the catalytic active site in toxic enzymes such as phospholipases A_2 and metalloproteinases. (b) Binding of antibodies to epitopes located close to the toxin active site, thus exerting inhibition by steric hindrance. (c) Binding of antibodies to molecular regions distant from the active/toxic site of venom components, neutralization being achieved by allosteric changes induced in the toxins, with the consequent reduction in their

ability to bind to tissue or cellular targets and to cause damage. (d) Formation of immunocomplexes between antibodies and toxins, with the subsequent removal by phagocytic cells; this last mechanism does not operate in the case of antivenoms made of monovalent Fab fragments, since they do not form complexes (Gutiérrez & León, 2009; Gutiérrez et al., 2011b).

6.3.1 Clinical performance of antivenoms: Efficacy

Antivenoms are administered parenterally, mostly by the intravenous route, and preferably diluted in physiological solution. Intradermal hypersentivity tests are not recommended since they have a very poor predictive value (Cupo et al., 1991; Malasit et al., 1986). The clinical performance of antivenoms depends on several factors associated with the immunological and physico-chemical characteristics of these products, as well as with the circumstances of their use in the clinical setting. At the preclinical level, antivenoms should be effective in the neutralization of the most relevant toxic activities of the venoms of medically-relevant snakes in a particular country or region. In some cases, this is achieved by using antivenoms raised against the venoms of the species that provoke the bite. In other cases, antivenoms are able to neutralize the venoms of species not used in the immunization of animals, but being phylogenetically related (WHO, 2010a). This phenomenon of immunological cross-reactivity has been clearly demonstrated, for instance, in the case of antivenoms raised against *Bothrops* sp venoms in Latin America (Otero et al., 1995; Segura et al., 2010a). In other cases, however, the cross-reactivity of antivenoms is low and, therefore, the efficacy of some products to neutralize venoms of medically-relevant species not included in immunization mixtures is limited, as occurs with venoms of some rattlesnakes and coral snakes in the Americas (Saravia et al., 2002; Tanaka et al., 2010). This issue of low cross-reactivity of some antivenoms may have potentially serious implications, when some products are used in the treatment of envenomings by species whose venoms are immunologically different from the ones used in immunization. One example has been the use of antivenoms manufactured in India for the treatment of envenomings in sub-Saharan Africa (Visser et al., 2008). This problem is complicated by the frequent lack of regulation and quality control of imported antivenoms in many countries, thus precluding the proper assessment of their neutralizing ability. This issue urges upgrading the regulatory capacities of countries in Asia, Africa and Latin America, as to ensure that antivenoms being introduced in these regions are evaluated with standard preclinical tests, such as those recommended by the WHO (2010a).

Antivenoms have demonstrated to be highly effective, when administered timely, at halting the most relevant systemic manifestations of snakebite envenoming (Gutiérrez & León, 2009; Lalloo & Theakston, 2003; Warrell, 1992). In the case of bites by viperids, systemic bleeding, hemodynamic manifestations and coagulation disturbances are controlled within hours after antivenom infusion. In contrast, toxins responsible for local pathological effects (edema, dermonecrosis, local hemorrhage and myonecrosis) are more difficult to neutralize by antivenoms, basically because the early onset of these effects upon venom injection, thus precluding an effective blockade by antivenom antibodies (Gutiérrez et al., 1998), a problem that is worsened by the occurrence of venom-induced vascular alterations, which affect the distribution of antivenom to the affected tissue (Battellino et al., 2003). In the case of neurotoxic venoms, characteristic of most elapid and some viperid species, the development of neurotoxic manifestations is prevented by the timely administration of antivenoms, with

the consequent neutralization of neurotoxins in the circulation before reaching neuromuscular junctions. However, neutralization is more difficult when neurotoxins are bound to receptors at the synapse. In the case of post-synaptic neurotoxins, their binding can be reverted (Alape-Girón et al., 1996; Boulain & Ménez, 1982), but presynaptically-acting toxins are known to destroy the nerve terminal, thus precluding neutralization and generating a more prolonged pattern of nerve damage (Prasarnpun et al., 2005). Thus, the clinical efficacy of antivenoms is intimately related to the ability of these products to bind with high affinity and neutralize relevant venom toxins located in tissues or in the bloodstream, as well as to the toxicokinetics of toxins and the pharmacokinetics of antivenom antibodies or antibody fragments (Gutiérrez et al., 2003; Scherrmann, 1994; WHO, 2010a). For instance, low molecular mass neurotoxins characteristic of elapid snake venoms are rapidly distributed and readily reach their targets in the neuromuscular junctions; in these cases, there is a mismatch between the toxicokinetics of these neurotoxins and the pharmacokinetics of antivenom antibodies (Gutiérrez et al., 2003; Ismail et al., 1998). On the other hand, low molecular mass Fab fragments have a relatively short half-life, thus resulting in the phenomenon of recurrence of envenoming, i.e. the reappearance of signs and symptoms of envenoming several hours after antivenom therapy (Ariaratnam et al., 1999; Boyer et al., 2001; Meyer et al., 1997). Careful clinical following up of patients is necessary to determine the need of an additional dose of antivenom.

The rapid access to effective antivenoms constitutes a key issue in the proper management of snakebite envenoming. If the envenoming is potentially severe, and if the access to antivenom is delayed, due to reasons that range from hesitation to use antivenoms to prolonged transportation times to health facilities and lack of antivenoms in health posts, the efficacy of antivenoms is jeopardized and various pathophysiological complications might ensue. Another factor that determines the efficacy of antivenom treatment has to do with the use of a correct dose of this immunobiological, and to the assessment of whether the patient needs an additional dose of antivenom, based on the evolution of clinical and laboratory parameters. These considerations demand that the health staff in charge of treating these envenomings have an adequate knowledge of the basic elements of antivenom usage.

6.3.2 Antivenom safety

Administration of antivenom is associated, in a variable percentage of cases, with early and late adverse reactions. Early adverse reactions (EARs) can be, in few cases, truly anaphylactic reactions, i.e. IgE-mediated, or, alternatively, anaphylactoid reactions, which occur more frequently, and are *de novo* reactions not mediated by previous exposure to horse proteins (Warrell, 1995a). The mechanisms of these reactions are not well understood, but are likely to depend on (a) complement activation by antibody aggregates present in antivenom (Sutherland, 1977); (b) formation of complexes between human heterophylic antibodies against antivenom antibodies, with consequent complement activation (León et al., 2008); or (c) presence of antibodies in antivenoms that react with cells, such as erythrocytes (León et al., 2007), leukocytes or endothelial cells, thus provoking adverse reactions. Such EARs can be mild, characterized by urticaria and itching only, or severe, involving angioedema, bronchospasm and hypotension (Warrell, 1995a). The incidence of EARs varies significantly among different antivenoms, from as low as 5% to higher than 70% of the cases with some products (Chippaux et al., 1998; Gawarammana et al., 2004;

Otero-Patiño et al., 1998). Such high variability is due to the different physicochemical quality of antivenoms, since some products have high protein concentration and high amounts of protein aggregates. Therefore, the physicochemical features of antivenoms greatly determine their safety profile, an issue that demands renewed efforts at the technological and regulatory levels. Another type of reaction observed in some antivenoms are pyrogenic reactions, associated with chills and fever (WHO, 2010b), but these should be avoided by a proper quality control, i.e. pyrogenicity testing, of these products. In the event of EARs, antivenom infusion should be stopped, and the reaction treated with adrenaline, anti-histamines and steroids (Warrell, 1995a). Once the reaction is controlled, antivenom infusion should be continued. Pretreatment with adrenaline has been advocated for reducing the incidence of EARs (de Silva et al., 2011). Late adverse reactions (LARs) to antivenoms occur 5-24 days after treatment, and are characterized by itching, fever, urticaria, arthralgia and proteinuria (Warrell, 1995a). This corresponds to a typical type III hypersensitivity reaction, i.e. serum sickness, due to the formation of immune complexes between antivenom antibodies and antibodies generated in the patient against antivenom proteins. The incidence of serum sickness after antivenom administration correlates with the amount of foreign protein, i.e. antivenom, administered (LoVecchio et al., 2003). LARs are treated with anti-histamines and steroids. Another aspect of antivenom safety that has to be considered is microbial safety, which is guaranteed by sterile filtration of the final product and the use of viral inactivation/removal steps (Burnouf et al., 2004; WHO, 2010a). Some of the manufacturing steps currently used in antivenom production inactivate or remove viruses, thus contributing to the microbial safety of these products (Burnouf et al., 2004; WHO, 2010a).

Such high heterogeneity in the safety of antivenoms, in terms of incidence of adverse reactions, calls for international cooperative efforts aimed at improving the technological platform of many antivenom producers, in order to increase the physicochemical quality of antivenoms on a world wide basis (Gutiérrez et al., 2011a). A number of antivenom producers in Asia, Africa and Latin America need to upgrade their facilities and protocols. The experience gained by well-developed antivenom manufacturing laboratories in various parts of the world should contribute to the improvement of less developed antivenom producers, through a variety of activities such as technology transfer programs, workshops, training and exchanges of various sorts. Such networking scenario should be promoted by the WHO and its regional offices, and by organizations such as the Global Snake Bite Initiative (www.snakebiteinitiative.org/).

6.4 Ancillary treatments

The therapy of snakebite envenoming includes a series of interventions in addition to antivenom administration. In the case of viperid venoms, hemodynamic and renal disturbances demand careful control of fluid therapy, monitoring of central venous pressure, and use of diuretics (Warrell, 1995a; WHO, 2010b). Infection often develops in viperid snakebites and requires the use of antibiotics. Moreover, local tissue damage by viperid and some elapid snakebites calls for debridement of necrotic tissue and care of the bitten limb. In some cases, when muscle intracompartmental pressures increase beyond 45 mm Hg, compartment syndrome ensues and fasciotomy is indicated (WHO, 2010b). In the case of neurotoxic envenomings caused by elapid and some viperid species, mechanical ventilation should be provided in the event of respiratory paralysis (Warrell, 1995a; WHO,

2010b). The complexity of snake venoms and the corresponding variability in the clinical presentation of these envenomings complicates the management of the cases and demands an adequate training of the health staff in charge of treating these emergencies, in order to guarantee the implementation of effective therapeutic interventions.

The poor efficacy of antivenoms to neutralize local tissue damage induced by viperid and some elapid venoms brings the need to find alternative therapies. A very promising avenue is the possibility of using natural or synthetic inhibitors of venom toxins, such as inhibitors of phospholipases A_2, metalloproteinases and hyaluronidases, for blocking the action of tissue-damaging toxins by rapidly administering these inhibitors directly on the site of venom injection (Gutiérrez et al., 2007; Lomonte et al., 2009; Perales et al., 2005). Such possibility has been tested, with excellent results, at the preclinical level in mouse models (Borkow et al., 1997; Lomonte et al., 2009; Rucavado et al., 2000; Yingprasertchai et al., 2003). It is necessary to identify and develop inhibitors, some of which may be already in use for other pathologies, and to test them at the preclinical and clinical levels. The therapy of snakebite envenoming in the future will likely involve, in addition to intravenous antivenom administration, the local injection of toxin inhibitors, as well as other ancillary interventions aimed at controlling the systemic aspects of envenoming and to modulate the deleterious aspects of the inflammatory response of the organism to snake venoms (Gutiérrez et al., 2007).

7. Preclinical and clinical testing of antivenoms

The large intra- and interspecies variability in the composition of snake venoms poses a problem for antivenom efficacy, since cross-neutralization of antivenoms against venoms not used in the immunizing mixture might not occur. Therefore, the distribution of antivenoms to countries or regions where medically-relevant snakes are different from those used in immunization schemes needs to be carefully evaluated in order to ensure that these antivenoms are indeed effective. This issue gets complicated by the fact that, quite often, regulatory agencies in developing countries do not have the facilities and expertise to perform adequate preclinical testing of the antivenoms being imported (D. Williams et al., 2011). A proper assessment of antivenom efficacy should be based on preclinical and clinical testing. At the preclinical level, it is necessary to assess the capacity of antivenoms to neutralize the lethal, as well as other relevant toxic activities, of the most important venoms in a country or region. This demands, in the first place, the establishment of local facilities to collect and keep medically-relevant snakes. These snake colonies should provide pools of venoms, which could then be used in preclinical testing of antivenoms. Precise indications on how to built and run these facilities are included in the *WHO Guidelines for Antivenom Production, Control and Regulation of Antivenoms* (WHO, 2010a). In the case of viperid venoms, a battery of preclinical tests usually includes the evaluation of the neutralization of lethal, hemorrhagic, coagulant, defibrinogenating and myotoxic activities (Gutiérrez et al., 2011b; Theakston, 1986; WHO, 2010a). In the case of elapid snakes, antivenom preclinical efficacy should be assessed by the neutralization of lethality and, in the case of elapid venoms that induce necrosis or coagulopathy, by the neutralization of dermonecrosis and coagulant activities, respectively (Gutiérrez et al., 2011b; WHO, 2010a). These methods involve simple laboratory procedures that need to be implemented in countries where antivenoms are being produced or imported. In addition, international collaborative

projects, involving well-developed laboratories, could be implemented in order to test antivenoms (D. Williams et al., 2011). More recently, a proteomic approach, named 'antivenomics', has been adapted for the evaluation of immune reactivity of antivenoms against particular toxins in venoms (Calvete, 2010; Gutiérrez et al., 2009a; Lomonte et al., 2008). This methodology allows for the identification of the toxins recognized by antivenom antibodies.

The preclinical assessment of antivenoms should be followed by clinical evaluation of antivenom safety and efficacy (WHO, 2010a). Since phase I clinical trials in healthy volunteers are ethically unacceptable in the case of antivenoms, because they might induce adverse reactions, a substitution of phase I clinical trial, by a protocol known as '3 + 3 dose escalation design', has been proposed for antivenoms (S.B. Abubakar et al., 2010b). This is then followed by phase III clinical trials in which a new antivenom is compared with an existing antivenom of known efficacy and safety (see for example the studies of I.S. Abubakar et al., 2010; Cardoso et al., 1993; Otero et al., 1999; Otero-Patiño et al., 1998; Smalligan et al., 2004; Warrell et al., 1974). Clinical trials should use robust clinical and laboratory end points for the assessment of therapeutic success. Furthermore, post-marketing surveillance (pharmacovigilance) is required to detect possible adverse reactions not reported in the clinical trials and to follow up efficacy (WHO, 2010a).

8. Technological aspects for antivenom improvement

The need to have antivenoms of wide cross-reactivity, able to neutralize venoms from as many snake species as possible, demands a careful revision of the design of venom mixtures used for immunization of animals. There is a large body of knowledge in the biochemistry, toxicology and immunology of snake venoms, especially of venoms from species having a heavy medical impact, which should be used for the re-evaluation of the immunizing mixtures and for the design of novel mixtures for new antivenoms (Gutiérrez et al., 2009a; D. Williams et al., 2011). Proteomics technologies, together with neutralization tests, constitute valuable tools to analyze venom composition and effects, and to assess the neutralizing profile of current and new antivenoms. The Global Snake Bite Initiative has proposed a strategy to structure an international collaborative effort to evaluate current antivenoms and to design improved antivenoms (D. Williams et al., 2011). One aspect of this strategy is based on the development of regional polyspecific antivenoms for use in sub-Saharan Africa and Asia using clinical, phylogenetic, proteomic and antivenomic analyses for the selection of the best venom mixtures for immunization. These antivenoms, manufactured by several laboratories, will then be evaluated by independent preclinical assessment, followed by clinical trials in various countries, performed by local medical personnel. In parallel, international expert committees will validate production facilities for prequalification, in a process aimed at ensuring the manufacture of the volume of antivenom needed in those regions (D. Williams et al., 2011).

One example of the potential usefulness of such an approach has to do with the design of immunizing mixtures for antivenoms to be used in sub-Saharan Africa. Several antivenoms use a mixture of venoms from many species; however, a recently developed antivenom was produced by using a mixture of venoms from only three species (Gutiérrez et al., 2005). Neutralization and antivenomic studies have shown that this new antivenom is able to

neutralize the venoms of several species of viperids and spitting cobras from sub-Saharan Africa (Calvete et al., 2010b; Segura et al., 2010b; Petras et al., 2011). Similarly, the ideal venom mixtures for antivenoms to be used in some parts of Asia need to be re-assessed on the basis of recent clinical evidence of the existence of medically-relevant species whose venoms are not routinely used in antivenom manufacture, such as that of the viperid *Hypnale hypnale* (Ariaratnam et al., 2008). Likewise, the decision on whether to prepare monospecific or polyspecific antivenoms has to be based on sound epidemiological, clinical, biochemical and immunological evidence. Consequently, the design and re-design of venom mixtures for immunization requires a multidisciplinary approach. On the other hand, there are other aspects of antivenom technological development that should be considered, such as stability and improved immunization schemes. Liquid antivenoms have to be stored at 2-8 ^{o}C (WHO, 2010a). However, the quality of the cold chain in many regions of the world is poor, thus complicating the distribution of antivenoms, especially to rural settings where most snakebites occur. This problem can be overcome by producing freeze-dried antivenoms, but this increases the production cost and, therefore, the price. Alternatives are being explored aimed at formulating liquid antivenoms stable at room temperature (Rodrigues-Silva et al., 1999; Segura et al., 2009). Likewise, the design of immunization protocols based on multi-site injection of small volumes containing low amounts of venom has resulted in higher neutralizing titers with very little damage to the immunized animals (Chotwiwatthanakun et al., 2001). Furthermore, the search for novel adjuvants is a relevant task in the efforts to improve antivenom antibody titers (Gutiérrez et al., 2011a).

9. The accessibility and correct use of antivenoms

Despite the widespread demonstration of antivenom efficacy for the treatment of snakebite envenoming, and the fact that many aspects of the know-how required to produce antivenoms are freely available (WHO, 2010a), there is a current deficit in antivenom accessibility in various regions of the world, most notably in sub-Saharan Africa and some countries of south-east Asia (Chippaux, 2010; Theakston et al., 2003; D. Williams et al., 2011; WHO, 2007a). This phenomenon has multiple causes, such as: (a) Withdrawal of some manufacturers from these markets due to profit considerations. (b) Privatization of former public laboratories, with the consequent increments in the prices of antivenoms. (c) The impact of international policies designed to reduce the size of the public sector, including a reduction in the provision of public health services and their privatization. (d) Weakening of antivenom manufacturing laboratories of the public sector in many developing countries, associated with lack of investment in facilities and technology, and reduction in training programs for the staff. (e) Lack of financial support for antivenom purchase by ministries of health, due to economic constraints and to prioritization on other health issues perceived as more pressing needs. (f) Loss of confidence in antivenom treatment in some regions due to the use of antivenoms of poor efficacy or safety. (g) Poor advocacy for promoting greater attention to snakebite envenoming as a neglected tropical disease. (h) Low profile of snakebite envenoming in the international public health agenda. As a result, antivenom accessibility is deficient in vast regions of Asia and Africa (WHO, 2007a; D. Williams et al., 2011). The solution to this complex problem demands concerted actions at various levels, from the technological and manufacturing realm to the public health arena (Chippaux, 2010; Gutiérrez et al., 2010b; D. Williams et al., 2010, 2011).

9.1 How to enhance the accessibility of antivenoms

Economic and political constraints constitute one of the main causes of poor accessibility of antivenoms in many countries. It is evident that the sole drive of the market forces will not solve this problem and, instead, well-designed strategies with a strong participation of governments and non-governmental organizations have to be implemented. This is a critical aspect that needs to be addressed by a variety of interventions such as: (a) Increasing the technological capacity of manufacturing laboratories in developing countries, both in the public and private realms, and introduction of cost-effective methodologies for antivenom production. One example is the manufacture of whole IgG antivenoms by caprylic acid fractionation of plasma (Gutiérrez et al., 2005; Rojas et al., 1994). This procedure generates antivenoms of high quality and high yield, at reduced production costs, thus constituting an excellent alternative for low-income countries (Brown & Landon, 2010). (b) Increased recognition of governments of low-income countries on the impact of snakebite envenoming as a public health problem, with the consequent political and financial decisions for the acquisition of adequate volumes of antivenom. (c) Using the capacity of large antivenom producers in order to manufacture antivenoms for other regions of the world at reasonable prices. This could be accomplished by promoting international partnerships between manufacturers, public health authorities, organizations of the civil society, and donors, similarly to what has been done for other neglected tropical diseases (Hotez et al., 2006). (d) Promoting strategies for price reduction, such as differential pricing arrangements or large scale 'pooled' purchases for various countries (Gutiérrez et al., 2010b).

9.2 Distribution of antivenoms: guaranteeing access to regions where snakebites occur

Even if governments purchase adequate volumes of antivenom, this does not guarantee that these drugs will reach the rural health posts where most snakebite envenomings occur. This problem has diverse roots, such as: (a) Incomplete information on the epidemiology of snakebites. In countries where official statistics of snakebite incidence are lacking or incomplete, the decision on where to distribute antivenoms cannot be taken on a rigorous base. This is another reason for underscoring the relevance of proper epidemiological register of this pathology. (b) Antivenom acquisition procedures by the ministries of health in many countries are slow and cumbersome; moreover, due to budgetary constraints, the volumes of antivenom purchased are often insufficient; both of these factors preclude the distribution of adequate volumes of this drug to rural settings. (c) Antivenoms are often distributed only to hospitals and clinics in large cities, distant from the regions where the majority of snakebites occur, thus affecting the timely treatment of patients. (d) As discussed previously, the lack of an adequate cold chain system in many rural settings of the world precludes the effective distribution of antivenoms. (e) Many rural regions are devoid of healthcare facilities, thus affecting the access to antivenom and other medical interventions and forcing people to travel large distances to receive medical attention.

This complex scenario demands the design of well-structured and effective strategies of antivenom distribution, on the basis of sound epidemiological information on snakebite incidence. An intersectorial and interprogrammatic approach should be promoted, in conjunction with other efforts being performed in the public health realm, in order to favour a synergy with other actors and projects, with the consequent impact in the cost-

effectiveness of interventions (WHO, 2007c). The compulsory report of snakebite envenoming (WHO, 2010a) and the use of geographical information systems to identify high risk areas (Hansson et al., 2010; Leynaud & Reati, 2009) would greatly contribute to generate a solid basis of information on the actual magnitude of the problem. Furthermore, the awareness of national and regional health authorities on the impact of snakebite envenoming should be promoted by academic, public health and civil society organizations, in order to ensure the acquisition and distribution of the required volumes of antivenom. Likewise, antivenom distribution strategies should benefit from the use of the cold chain system already developed for vaccine distribution. Also, the provision of antivenom access to rural settings and the training of rural health staff in the correct administration of antivenoms should be prioritized. Interventions tailored to the conditions of each country and region should be promoted, in order to optimize the available resources and guarantee a rapid access to treatment (see for example Otero et al., 1992).

9.3 Promoting the correct use of antivenoms

The distribution of antivenoms to regions where snakebites occur should be complemented by a proper training of health staff in the correct usage of this product and in the proper treatment of snakebite envenomings. There is evidence of poor knowledge of medical and nursing staff in various regions of the world on how to diagnose and treat snakebite envenomings, how to use antivenoms, and how to treat possible adverse reactions to their administration (Gutiérrez et al., 2009c; Simpson, 2008). This requires concerted efforts at medical and nursing schools in the universities, as well as the implementation of permanent educational programs on this subject, particularly aimed at rural health facilities. Likewise, the implementation of teaching material and the development of guidelines for the diagnosis and treatment of snakebite envenomings should be actively promoted, both at regional (WHO, 2010b) and national levels. These tasks should involve not only teaching institutions, but also public health authorities, local organizations of the civil society, and antivenom manufacturers. The critical revision of antivenom prospects, on the basis of current knowledge on the taxonomy of snakes and on the clinics of snakebite envenomings, are of great relevance, in the light of evident misconceptions included in the prospects of some antivenoms (Simpson & Norris, 2007).

10. Prevention of snakebites

Prevention programs aimed at reducing the impact and incidence of snakebite envenomings should be a priority in the international efforts required to confront this problem. The design of these programs should be tailored to the cultural, social, economic and institutional characteristics of the populations, and should involve the active participation of the communities in their design and implementation. Impoverished and excluded groups, such as indigenous communities in many parts of the world, should receive particular attention. The design of these programs should be also based on sound social science research aimed at understanding the particularities and needs of each region and context, with the participation of the communities. It is highly relevant, for instance, to understand how the problem is perceived in the community and what types of preventive interventions are suited for each particular context. Likewise, specific strategies should be designed for situations involving natural disasters, as snakebites have been reported to increase in such

circumstances. In addition, the natural history of envenomings should be considered, including the distribution of snakes in various types of crops and the behaviour of snakes. In some regions of Asia, bites by kraits (genus *Bungarus*) often occur at nights inside human dwellings while people are asleep on the ground (Sharma et al., 2004). The use of mosquito nets has reduced the incidence of envenoming by kraits in Nepal (Chapuis et al., 2007). The majority of viperid snakebites occur in the feet; thus, a preventive measure should be the use of footware (Alirol et al., 2010; Gutiérrez, 2010; Warrell, 2010).

11. Final remarks: The need for an integrated approach and for the promotion of partnerships

The world wide efforts required to reduce the impact of snakebite envenoming should be conceptualized within the frame of the Millennium Development Goals (MDGs) (http://www.un.org/millenniumgoals/global.shtml), particularly regarding the provision of access to essential drugs (WHO, 2011), in this case antivenoms, to developing countries. The access to adequate health services is a human right, and states and other international instances have the obligation to ensure the access to health facilities, goods and services on a non-discriminatory basis, especially to vulnerable and marginalized groups, and to provide education and access to information to the communities on relevant health issues, such as snakebite envenoming. Therefore, interventions aimed at ameliorating the impact of this pathology should be viewed within a frame of human rights and social responsibility of states, international organizations and non-governmental groups.

Snakebite envenoming is a 'tool-ready' disease, in the sense that the basic technological therapeutic tools to treat this pathology, i.e. antivenoms, are available. However, there is a need to implement renewed efforts to improve the quality of some antivenoms, to design new antivenoms for various regions in the world, and to increase the volume of production as to fulfil the world wide needs to these immunobiologicals. Scientific, technological and public health tasks include acquisition of more rigorous data on the incidence of snakebite incidence and mortality, assessment of preclinical and clinical performance of currently available antivenoms, and development of novel antivenoms on the basis of epidemiological, biochemical, toxicological and immunological knowledge on venoms. Moreover, the strengthening of antivenom manufacture on a global basis should involve an active process of technology transfer and training aimed at improving the current technological platform of many antivenom producers, especially those located in developing countries. Finally, renewed efforts should be undertaken to guarantee the deployment and effective distribution and use of antivenoms to the regions of the world where this pathology has its highest impact. Table 1 summarizes some of the most pressing tasks that need to be promoted as part of a global strategy to reduce the impact of snakebite envenomings.

The design of effective strategies to confront this problem should be also integrated with the more general efforts in the arena of neglected tropical diseases (WHO, 2007c). Such strategies should be conceived from an intersectorial and interprogrammatic perspective (see WHO, 2007c), with a synergistic approach involving the control of other neglected tropical diseases; such an approach will significantly increase the cost-effectiveness of interventions. Areas of possible interprogrammatic cooperation include: (a) The collaborative delivery of antivenoms

within distribution channels already developed for other immunobiologicals, such as vaccines. (b) Incorporating antivenoms in integrated drug purchasing schemes in developing countries on a regional basis. (c) Strengthening the development of public health systems in remote rural areas where snakebite envenomings are frequent. (d) Promoting partnerships of diverse sorts with groups involved in the combat of neglected tropical diseases, such as foundations, non-governmental organizations and other advocacy groups. (e) Including snakebite envenoming in the agenda of organizations that provide financial support for research and intervention programs for neglected tropical diseases in developing countries. (f) Incorporating the subject of snakebite prevention, diagnosis and treatment within the context of educational packages on neglected tropical diseases in teaching institutions, public health facilities and communities. (g) Promoting the inclusion of the subject of snakebite envenoming within the agenda of community organizations for the promotion of health in rural areas of countries in Africa, Asia and Latin America.

1. Acquisition of reliable information on snakebite incidence and mortality
2. Innovation in the technology for the production of antivenoms
3. Strengthening the capacity of laboratories in low-income countries to manufacture and control antivenoms
4. Commitment of manufacturers to produce antivenoms for regions devoid of local production
5. Implementation of economic strategies to ensure the sustainable production of antivenoms
6. Improvement of the national regulatory expertise and quality control of antivenoms in low-income countries
7. Accessibility of antivenoms at affordable prices in low-income countries
8. Preclinical and clinical assessment of antivenom efficacy and safety
9. Development of effective antivenom distribution programs to regions of high incidence of snakebites
10. Permanent training programs for health staff on snakebite envenomings and their treatment
11. Development of programs to support people suffering from sequelae of snakebite envenomings
12. Preventive and educational programs at the community level with involvement of local organizations

Table 1. Summary of some of the most important tasks for an integrated strategy to confront the problem of snakebite envenoming from a global perspective. Adapted from Gutiérrez et al. (2010b)

In the long term, the reduction of the impact of snakebite morbidity and mortality, with its drastic effects on the quality of human life, should involve a global partnership incorporating many different actors at various levels in our societies, such as: (a) The scientific ('epistemic') community of toxinologists, represented by the International Society on Toxinology (IST) and researchers in every continent. (b) Groups working on technological development and technology transfer activities, both in the pharmaceutical industry and in university departments. (c) Antivenom manufacturers. (d) International health organizations, especially the WHO and its regional offices. (e) National public health

authorities, i.e. Ministries of Health and other organizations of the public health sector. (f) National regulatory bodies, responsible for ensuring the safety and efficacy of antivenoms being distributed. (g) Non-governmental organizations (NGOs) and advocacy groups that promote a public health agenda and the access of essential drugs to developing countries. (h) Organizations of the civil society of countries having a high burden of snakebite envenoming (Figure 2). The current tasks of generating a growing international awareness on the magnitude of this problem, establishing partnerships to ensure the development, availability and accessibility to antivenoms, and promoting prevention and an effective clinical management of this pathology, are being promoted by the Global Snake Bite Initiative (GSI), the WHO, and a number of national and regional projects in various parts of the world.

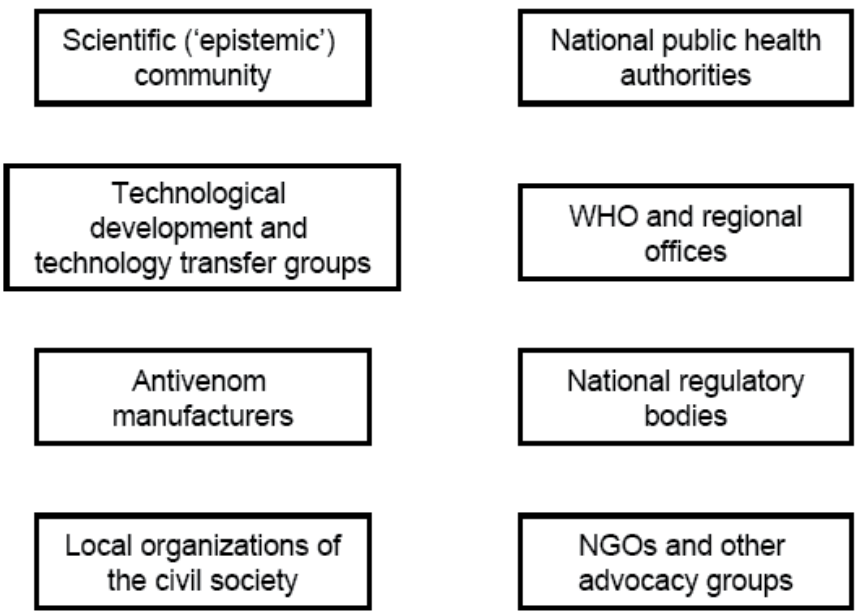


Fig. 2. Some of the participants that should be involved in a global partnership aimed at the reduction of the impact of snakebite envenoming in the world.

12. Acknowledgements

The author thanks his colleagues of Instituto Clodomiro Picado, University of Costa Rica, and of other groups in Latin America and elsewhere, for long-standing collaborations in this subject. Special thanks are due to David A. Warrell, David Williams, Fan Hui Wen, João Luiz Costa Cardoso, Rafael Otero-Patiño, Robert Harrison, Kenneth D. Winkel, R. David G. Theakston, Juan José Calvete, Jean Philippe Chippaux, Abdusalam Nasidi, Abdulrazaq Habib, Ulrich Kuch, and Thierry Burnouf for valuable discussions and cooperation in the field of snakebite envenoming and treatment. David Williams and Mark O'Shea kindly provided photographs of snakes. Many of the studies cited in this review have been

supported by Vicerrectoría de Investigación (University of Costa Rica), the International Foundation for Science (IFS), CRUSA-CSIC, NeTropica and the program CYTED.

13. References

Abubakar, I.S.; Abubakar, S.B.; Habib, A.G.; Nasidi, A.; Durfa, N.; Yusuf, P.O.; Larnyang, S.; Garnvwa, J.; Sokomba, E.; Salako, L.; Theakston, R.D.G.; Juszczak, E.; Alder, N. & Warrell, D.A. (2010). Randomised controlled double-blind non-inferiority trial of two antivenoms for saw-scaled or carpet viper (*Echis ocellatus*) envenoming in Nigeria. *PLoS Neglected Tropical Diseases*, Vol.4, No.7, pp. e767

Abubakar, S.B.; Habib, A.G. & Mathew, J. (2010a). Amputation and disability following snakebite in Nigeria. *Tropical Doctor*, Vol.40, No.2, pp. 114-116

Abubakar, S.B.; Abubakar, I.S.; Habib, A.G.; Nasidi, A.; Durfa, N.; Yusuf, P.O.; Larnyang, S.; Garnvwa, J.; Sokomba, E.; Salako, L.; Laing, G.D.; Theakston, R.D.G.; Juszczak, E.; Alder, N. & Warrell, D.A. (2010b). Pre-clinical and preliminary dose-finding and safety studies to identify candidate antivenoms for treatment of envenoming by saw-scaled or carpet vipers (*Echis ocellatus*) in northern Nigeria. *Toxicon*, Vol.55, No.4, pp. 719-723

Alape-Girón, A.; Stiles, B.G. & Gutiérrez, J.M. (1996). Antibody-mediated neutralization and binding-reversal studies on α-neurotoxins from *Micrurus nigrocinctus nigrocinctus* (coral snake) venom. *Toxicon*, Vol.34, No.3, pp. 369-380

Alape-Girón, A.; Sanz, L.; Escolano, J.; Flores-Díaz, M.; Madrigal, M.; Sasa, M. & Calvete, J.J. (2008). Snake venomics of the lancehead pitviper *Bothrops asper*: geographic, individual, and ontogenetic variations. *Journal of Proteome Research*, Vol.7, No.8, pp. 3556-3571

Alirol, E.; Sharma, S.K.; Bawaskar, H.S.; Kuch, U. & Chappuis, F. (2010). Snake bite in South Asia: A review. *PLoS Neglected Tropical Diseases*, Vol.4, No.1, pp. e603

Ariaratnam, C.A.; Meyer, W.P.; Perera, G.; Addleston, M.; Kuleratne, S.A.; Attapattu, W.; Sheriff, R.; Richards, A.M.; Theakston, R.D.G. & warrell, D.A. (1999). A new monospecific ovine Fab fragment antivenom for treatment of envenoming by the Sri Lankan Russell's viper (*Daboia russelli russelli*): a preliminary dose-finding and pharmacokinetic study. *American Journal of Tropical Medicine and Hygiene*, Vol. 61, No.2, pp. 259-265

Ariaratnam, C.A.; Thuraisingam, V.; Kularatne, S.A.; Sheriff, M.H.; Theakston, R.D.G.; de Silva, A. & Warrell, D.A. (2008). Frequent and potentially fatal envenoming by hump-nosed pit vipers (*Hypnale hypnale* and *H. nepa*) in Sri Lanka: lack of effective antivenom. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol.102, No.11, pp. 1120-1126

Ariaratnam, C.A.; Sheriff, M.H.; Arambepola, C.; Theakston, R.D.G. & Warrell, D.A. (2009). Syndromic approach to treatment of snake bite in Sri Lanka based on results of a prospective national hospital-based survey of patients envenomed by identified snakes. *American Journal of Tropical Medicine and Hygiene*, Vol.81, No.4, pp. 725-731

Azevedo-Marques, M.M.; Hering, S.E. & Cupo, P. (2009). Acidente crotálico. In: *Animais Peçonhentos no Brasil. Biologia, Clínica e Terapêutica dos Acidentes*, *2nd. Edition*, J.L.C. Cardoso; F.O.S. França; H.W. Fan; C.M.S. Málaque & V. Haddad, (Eds), Sarvier, 108-115, ISBN 978-85-7378-194-6, São Paulo, Brazil.

Battellino, C.; Piazza, R.; da Silva, A.M.; Cury, Y. & Farsky, S.H.P. (2003). Assessment of the efficacy of bothropic antivenom therapy on microcirculatory effects induced by *Bothrops jararaca* snake venom. *Toxicon*, Vol.41, No.5, pp. 583-593

Bdolah, A. (2010). Sarafotoxins, the snake venom homologs of the endothelins, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 303-315, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Bon, C. (1996). Serum therapy was discovered 100 years ago. In: *Envenomings and Their Treatments*, C. Bon & M. Goyffon, (Eds), 3-9, Fondation Marcel Mérieux, Lyon, France.

Bon, C. (1997). Multicomponent neurotoxic phospholipases A_2. In: *Venom Phospholipase A_2 Enzymes: Structure, Function and Mechanism*, R.M. Kini, (Ed.), 269-285, Wiley, ISBN 0-471-96189-2, Chichester, United Kingdom.

Borkow, G.; Gutiérrez, J.M. & Ovadia, M. (1997). Inhibition of toxic activities of *Bothrops asper* venom and other crotalid snake venoms by a novel neutralizing mixture. *Toxicology and Applied Pharmacology*, Vol.147, No.2, pp. 442-447

Boulain, J.C.; Ménez, A. (1982). Neurotoxin-specific immunoglobulins accelerate dissociation of the neurotoxin-acetylcholine receptor complex. *Science*, Vol.217, pp. 732-733

Boyer, L.V.; Seifert, S.A. & Cain, J.S. (2001). Recurrence phenomena after immunoglobulin therapy for snake envenomations: Part 2. Guidelines for clinical management with crotaline Fab antivenom. *Annals of Emergency Medicine*, Vol.37, No.2, pp. 196-201

Brown, N. & Landon, J. (2010). Antivenom: the most cost-effective treatment in the world? *Toxicon*, Vol.55, No.7, pp. 1405-1407

Burnouf, T.; Griffiths, E.; Padilla, A.; Seddick, S.; Stephano, M.A. & Gutiérrez, J.M. (2004). Assessment of the viral safety of antivenoms fractionated from equine plasma. *Biologicals*, Vol.32, No.3, pp. 115-128

Calvete, J.J.; Juárez, P. & Sanz, L. (2007). Snake venomics. Strategy and applications. *Journal of Mass Spectrometry*, Vol.42, No.11, pp. 1405-1414

Calvete, J.J.; Sanz, L.; Angulo, Y.; Lomonte, B. & Gutiérrez, J.M. (2009). Venoms, venomics, antivenomics. *FEBS Letters*, Vol.583, No.11, pp. 1736-1743

Calvete, J.J. (2010). Antivenomics and venom phenotyping: A marriage of convenience to address the performance and range of clinical use of antivenoms. *Toxicon*, Vol.56, No.7, pp. 1284-1291

Calvete, J.J.; Sanz, L.; Cid, P.; de la Torre, P.; Flores-Díaz, M.; Dos Santos, M.C.; Borges, A.; Bremo, A.; Angulo, Y.; Lomonte, B.; Alape-Girón, A. & Gutiérrez, J.M. (2010a). Snake venomics of the Central American rattlesnake *Crotalus simus* and the South American *Crotalus durissus* complex points to neurotoxicity as an adaptive paedomorphic trend along Crotalus dispersal in South America. *Journal of Proteome Research*, Vol.9, No.1, pp. 528-544

Calvete, J.J.; Cid, P.; Sanz, L.; Segura, A.; Villalta, M.; Herrera, M.; León, G.; Harrison, R.; Durfa, N.; Nasidi, A.; Theakston, R.D.G.; Warrell, D.A. & Gutiérrez, J.M. (2010b). Antivenomic assessment of the immunological reactivity of EchiTAb-Plus-ICP, an antivenom for the treatment of snakebite envenoming in sub-Saharan Africa. *American Journal of Tropical Medicine and Hygiene*, Vol.82, No.6, pp 1194-1201

Cardoso, J.L.C.; Fan, H.W.; França, F.O.S.; Jorge, M.T.; Leite, R.P.; Nishioka, S.A.; Avila, A.; Sano-Martins, I.S.; Tomy, S.C.; Santoro, M.L.; Chudzinski, A.M.; Castro, S.C.B.; Kamiguti, A.S.; Kelen, E.M.A.; Hirata, M.H.; Mirandola, R.M.S.; Theakston, R.D.G. & Warrell, D.A. (1993). Randomized comparative trial of three antivenoms in the treatment of envenoming by lance-headed vipers (*Bothrops jararaca*) in São Paulo, Brazil. *Quarterly Journal of Medicine*, Vol.86, No.5, pp. 315-325

Casewell, N.R.; Wagstaff, S.C.; Harrison, R.A.; Renjifo, C. & Wüster, W. (2011). Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom

metalloproteinase toxin genes. *Molecular Biology and Evolution*, Vol.28, No.9, pp. 2637-2649.

Chappuis, F.; Sharma, S.K.; Jha, N.; Loutan, L. & Bovier, P.A. (2007). Protection against snake bites by sleeping under a bed net in southeastern Nepal. *American Journal of Tropical Medicine and Hygiene*, Vol.77, No.1, pp. 197-199

Chippaux, J.P.; Williams, V. & White, J. (1991). Snake venom variability: methods of study, results and interpretation. *Toxicon*, Vol.29, No.11, pp. 1279-1303

Chippaux, J.P. (1998). Snake-bites: appraisal of the global situation. *Bulletin of the World Health Organization*, Vol.76, No.5, pp. 515-524

Chippaux, J.P.; Lang, J.; Eddine, S.A.; Fagot, P.; Rage, V.; Peyrieux, J.C. & Le Mener, V. (1998). Clinical safety of a polyvalent $F(ab')_2$ equine antivenom in 223 African snake envenomations: a field trial in Cameroon. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol.92, No.6, pp. 657-662

Chippaux, J.P. (2010). Snakebite in Africa. Current situation and urgent needs, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 453-473, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Chippaux, J.P. (2011). Estimate of the burden of snakebites in sub-Saharan Africa: a meta-analytic approach. *Toxicon*, Vol.57, No.4, pp. 586-599

Chotwiwatthanakun, C.; Pratanaphon, R.; Akesowan, S.; Sriprapat, S. & Ratanabanangkoon, K. (2001). Production of potent polyvalent antivenom against three elapid venoms using a low dose, low volume, multi-site immunization protocol. *Toxicon*, Vol.39, No.10, pp. 1487-1494

Corrêa-Netto, C.; Junqueira-de-Azevedo, I.; Silva, D.A.; Ho, P.L.; Leitão-de-Araújo, M.; Alves, M.L.; Sanz, L.; Foguel, D.; Zingali, R.B. & Calvete, J.J. (2011). Snake venomics and venom gland transcriptomic analysis of Brazilian coral snakes, *Micrurus altirostris* and *M. corallinus*. *Journal of Proteomics*, Vol.74, No.9, pp. 1795-1809

Cupo, P.; Azevedo-Marques, M.M.; de Menezes, J.B. & Hering, S.E. (1991). Reações de hipersensibilidade imediatas após uso intravenoso de soros antivenenos: valor prognóstico dos testes de sensibilidade intradérmicos. *Revista Instituto de Medicina Tropical de São Paulo*, Vol.33, No.2, pp. 115-122

Dart, R.C.; Mcnally, J.T.; Spaite, D.W. & Gustafson, R. (1992). The sequelae of pitviper poisoning in the United States. In: Biology of the Pitvipers, J.A. Cambpell & E.D. Brodie, (Eds.), 395-404, Selva, ISBN 0-9630537-0-1, Texas, USA.

de Oliveira, R.C.; Fan, H.W. & Sifuentes, D.N. (2009). Epidemiologia dos accidentes por animais peçonhentos. In: *Animais Peçonhentos no Brasil. Biologia, Clínica e Terapêutica dos Acidentes*, *2nd Edition*, J.L.C. Cardoso; F.O.S. França; H.W. Fan; C.M.S. Málaque & V. Haddad, (Eds), Sarvier, 6-21, ISBN 978-85-7378-194-6, São Paulo, Brazil.

de Silva, H.A.; Pathmeswaran, A.; Ranasinha, C.D.; Jayamanne, S.; Samarakoon, S.B.; Hittharage, A.; Kalupahana, R.; Ratnatilaka, G.A.; Uluwatthage, W.; Aronson, J.K.; Armitage, J.M.; Lalloo, D.G & de Silva, H.J. (2011). Low-dose adrenaline, promethazine, and hydrocortisone in the prevention of acute adverse reactions to antivenom following snakebite: a randomised, double-blind, placebo-controlled trial. *PLoS Medicine*, Vol.8, No.5, pp. e1000435

dos Santos, M.C.; D'Imperio-Lima, M.R.; Furtado, G.C.; Colletto, G.M.; Kipnis, T.L & Dias da Silva, W. (1989). Purification of $F(ab')_2$ anti-snake venom by caprylic acid: a fast method for obtaining IgG fragments with high neutralization activity, purity and yield. *Toxicon*, Vol.27, No.3, pp. 297-303.

Dufton, M.J. & Hider, R.C. (1988). Structure and pharmacology of elapid cytotoxins. *Pharmacology and Therapy*, Vol.36, No.1, pp. 1-40

Escalante, T.; Rucavado, A.; Fox, J.W. & Gutiérrez, J.M. (2011). Key events in microvascular damage induced by snake venom hemorrhagic metalloproteinases. *Journal of Proteomics*, Vol.74, No.9, pp. 1781-1794

Faiz, A.; Ghose, A.; Ahsan, F.; Rahman, R.; Amin, R.; Hassan, M.U.; Chowdhury, A.W.; Kuch, U.; Rocha, T.; Harris, J.B.; Theakston, R.D.G. & Warrell, D.A. (2010). The greater black krait (*Bungarus niger*), a newly recognized cause of neuro-myotoxic snake bite envenoming in Bangladesh. *Brain*, Vol.133, No.11, pp. 3181-3193

Fan, H.W. & Cardoso J.L. (1995). Clinical toxicology of snake bites in South America, In: *Handbook of Clinical Toxicology of Animal Venoms and Poisons*, J. Meier & J. White, (Ed.), 667-688, CRC Press, ISBN 0-8493-4489-1, Boca Raton, USA

Ferquel, E.; de Haro, L.; Jan, V.; Guillemin, I.; Jourdain, S.; Teynié, A.; d'Alayer, J. & Choumet, V. (2007). Reappraisal of *Vipera aspis* venom neurotoxicity. *PLoS ONE*, Vol.2, No.11, pp. e1194

Fox, J.W. & Serrano, S.M.T. (2005). Structural considerations of the snake venom metalloproteinases, key members of the M12 reprolysin family of metalloproteinases. *Toxicon*, Vol.45, No.8, pp. 969-985

Fox, J.W. & Serrano S.M.T. (2008). Exploring snake venom proteomes: multifaceted analyses for complex toxin mixtures. *Proteomics*, Vol.8, No.4, pp. 909-920

Fry, B.G.; Vidal, N.; Norman, J.A.; Vonk, F.J.; Scheib, H.; Ramjan, S.F.; Kuruppu, S.; Fung, K.; Hedges, S.B.; Richardson, M.K.; Hodgson, W.C.; Ignjativoc, V.; Summerhayes, R. & Kochva, E. (2006). Early evolution of the venom system in lizards and snakes. *Nature*, Vol.439, pp. 584-588

Fry, B.G.; Vidal, N.; van der Weerd, L.; Kochva, E. & Renjifo, C. (2009) Evolution and diversification of the Toxicofera reptile venom system. *Journal of Proteomics*, Vol.72, No.2, pp. 127-136

Gawarammana, I.B.; Kularatne, S.A.; Dissanayake, W.P.; Kumarasiri, R.P.; Senanayake, N. & Ariyasena, H. (2004). Parallel infusion of hydrocortisone +/- chlorpheniramine bolus injection to prevent acute adverse reactions to antivenom for snakebites. *Medical Journal of Australia*, Vol.180, No.1, pp. 20-23

Gómez, H.F. & Dart, R.C. (1995). Clinical toxicology of snakebite in North America, In: *Handbook of Clinical Toxicology of Animal Venoms and Poisons*, J. Meier & J. White, (Eds.), 619-644, CRC Press, ISBN 0-8493-4489-1, Boca Raton, USA

Grandgeorge, M.; Véron, J.L.; Lutsch, C.; Makula, M.F.; Riffard, P.; Pépin, S. & Scherrmann, J.M. (1996). Preparation of improved $F(ab')_2$ antivenoms. An example: new polyvalent European viper antivenom (equine). In: *Envenomings and Their Treatments*, C. Bon & M. Goyffon, (Eds), 161-172, Fondation Marcel Mérieux, Lyon, France

Gutiérrez, J.M.; León, G.; Rojas, G.; Lomonte, B.; Rucavado, A. & Chaves, F. (1998). Neutralization of local tissue damage induced by *Bothrops asper* (terciopelo) snake venom. *Toxicon*, Vol.36, No.11, pp. 1529-1538

Gutiérrez, J.M. & Ownby, C.L. (2003). Skeletal muscle degeneration induced by venom phospholipases A_2: insights into the mechanisms of local and systemic myotoxicity. *Toxicon*, Vol.42, No.8, pp. 915-931

Gutiérrez, J.M.; León, G. & Lomonte, B. (2003). Pharmacokinetic-pharmacodynamic relationships of immunoglobulin therapy for envenomation. *Clinical Pharmacokinetics*, Vol.42, No.8, pp. 721-741.

Gutiérrez, J.M.; Rojas, E.; Quesada, L.; León, G.; Núñez, J.; Laing, G.D.; Sasa, M.; Renjifo, J.M.; Nasidi, A.; Warrell, D.A.; Theakston, R.D.G. & Rojas, G. (2005). Pan-African polyspecific antivenom produced by caprylic acid purification of horse IgG: an

alternative to the antivenom crisis in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol. 99, No.6, pp. 468-475

Gutiérrez, J.M.; Theakston, R.D.G. & Warrell, D.A. (2006). Confronting the neglected problem of snake bite envenoming: the need for a global partnership. *PLoS Medicine*, Vol. 3, No.6, pp. e150

Gutiérrez, J.M.; Lomonte, B.; León, G.; Rucavado, A.; Chaves, F. & Angulo, Y. (2007). Trends in snakebite envenomation therapy: scientific, technological and public health considerations. *Current Pharmaceutical Design*, Vol.13, No.28, pp. 2935-2950

Gutiérrez, J.M. & León, G. (2009). Snake antivenoms. Technological, clinical and public health issues. In: *Animal Toxins: State of the Art. Perspectives in Health and Biotechnology*, M.E. de Lima; A.M.C. Pimenta; M.F. Martin-Euclaire; R.B. Zingali & H. Rochat, (Eds.), 393-421, Editora UFMG, ISBN 978-85-7041-735-0, Belo Horizonte, Brazil.

Gutiérrez, J.M. & Lomonte, B. (2009). Efectos locales en el envenenamiento ofídico en América Latina. In: *Animais Peçonhentos no Brasil. Biologia, Clínica e Terapêutica dos Acidentes*, J.L.C. Cardoso; F.O.S. França; H.W. Fan; C.M.S. Málaque & V. Haddad, (Eds), Sarvier, 352-365, ISBN 978-85-7378-194-6, São Paulo, Brazil.

Gutiérrez, J.M.; Lomonte, B.; León, G.; Alape-Girón, A.; Flores-Díaz, M.; Sanz, L.; Angulo, Y. & Calvete, J.J. (2009a). Snake venomics and antivenomics: Proteomic tools in the design and control of antivenoms for the treatment of snakebite envenoming. *Journal of Proteomics*, Vol.72, No.2, pp. 165-182

Gutiérrez, J.M.; Escalante, T. & Rucavado, A. (2009b). Experimental pathophysiology of systemic alterations induced by *Bothrops asper* snake venom. *Toxicon*, Vol.54, No.7, pp. 976-987

Gutiérrez, J.M.; Fan, H.W.; Silvera, C.L. & Angulo, Y. (2009c). Stability, distribution and use of antivenoms for snakebite envenomation in Latin America: report of a workshop. *Toxicon*, Vol.53, No.6, pp. 625-630

Gutiérrez, J.M. (2010). Snakebite envenomation in Central America, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 491-507, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Gutiérrez, J.M.; Rucavado, A. & Escalante, T. (2010a). Snake venom metalloproteinases. Biological roles and participation in the pathophysiology of envenomation, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 115-138, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Gutiérrez, J.M.; Williams, D.; Fan, H.W. & Warrell, D.A. (2010b). Snakebite envenoming from a global perspective: Towards an integrated approach. *Toxicon*, Vol.56, No.7, pp. 1223-1235

Gutiérrez, J.M. (2011). Envenenamientos por mordeduras de serpientes en América Latina y el Caribe: Una visión integral de carácter regional. *Boletín de Malariología y Salud Ambiental*, Vol.51, No.1, pp. 1-16

Gutiérrez, J.M.; León, G. & Burnouf, T. (2011a). Antivenoms for the treatment of snakebite envenomings: the road ahead. *Biologicals*, Vol.39, No.3, pp. 129-142

Gutiérrez, J.M.; León, G.; Lomonte, B. & Angulo, Y. (2011b). Antivenoms for snakebite envenomings. *Inflammation & Allergy-Drug Targets*, Vol. 10, No.5, pp. 369-380

Habib, A.G.; Gebi, U.I. & Onyemelukwe, G.C. (2001). Snake bite in Nigeria. *African Journal of Medicine and Medical Sciences*, Vol.30, pp. 171-178

Hansson, E.; Cuadra, S.; Oudin, A.; de Jong, K.; Stroh, E.; Torén, K. & Albin, M. (2010). Mapping snakebite epidemiology in Nicaragua-Pitfalls and possible solutions. *PLoS Neglected Tropical Diseases*, Vol.4, No.11, pp. e896

Hardy, D.L. (2009). Alternatives in the field management of venomous snakebite. In: *Animais Peçonhentos no Brasil. Biologia, Clínica e Terapêutica dos Acidentes*, 2*nd Edition,* J.L.C. Cardoso; F.O.S. França; H.W. Fan; C.M.S. Málaque & V. Haddad, (Eds), Sarvier, 454-468, ISBN 978-85-7378-194-6, São Paulo, Brazil.

Harrison, R.A.; Hargreaves, A.; Wagstaff, S.C.; Faragher, B. & Lalloo, D.G. (2009). Snakebite envenoming: a disease of poverty. *PLoS Neglected Tropical Diseases*, Vol.3, No.12, pp. e569

Harvey, A.L. (2001). Twenty years of dendrotoxins. *Toxicon*, Vol.39, No.1, pp. 15-26

Harvey, A.L. (2010). Fasciculins. Toxins from mamba venoms that inhibit acetylcholinesterase, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 317-324, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Hegde, R.P.; Rajagopalan, N.; Doley, R. & Kini, R.M. (2010). Snake venom three-finger toxins, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 287-301, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Hotez, P.J.; Molyneux, D.H.; Fenwick, A.; Ottesen, E.; Ehrlich-Sachs, S. & Sachs, J.D. (2006). Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. *PLoS Medicine*, Vol.3, No.5, pp. e102

Ismail, M.; Abd-Elsalam, M.A. & Al-Ahaidib, M.S. (1998). Pharmacokinetics of ^{125}I-labelled *Walterinnesia aegyptia* venom and its specific antivenins: flash absorption and distribution of the venom and its toxin versus slow absorption and distribution of IgG, $F(ab')_2$ and Fab of the antivenin. *Toxicon*, Vol.36, No.1, pp. 93-114

Jayanthi G.P. & Gowda, T.V. (1988). Geographical variation in India in the composition and lethal potency of Russell's viper (*Vipera russelli*) venom. *Toxicon*, Vol.26, No.3, pp. 257-264

Kasturiratne, A.; Wickremasinghe, A.R.; de Silva, N.; Gunawardena, N.K.; Pathmeswaran, A.; Premaratna, R.; Savioli, L.; Lalloo, D.G. & de Silva, H.J. (2008). The global burden of snakebite: a literature analysis and modeling based on regional estimates of envenoming and deaths. *PLoS Medicine*, Vol.5, No.11, pp. e218

Kini R.M & Chan, Y.M. (1999). Accelerated evolution and molecular surface of venom phospholipase A_2 enzymes. *Molecular Evolution*, Vol.48, No.2, pp. 125-132

Kini, R.M. (2005). The intriguing world of prothrombin activators from snake venom. *Toxicon*, Vol.45, No.8, pp. 1133-1145

Kulkeaw, K.; Chaicumpa, W.; Sakolvaree, Y.; Tongtawe, P. & Tapchaisiri, P. (2007). Proteome and immunome of the venom of the Thai cobra, *Naja kaouthia. Toxicon*, Vol.49, No.7, pp. 1026-104

Lalloo, D.G. & Theakston, R.D.G. (2003). Snake antivenoms. *Journal of Toxicology-Clinical Toxicology*, Vol. 41, No.3, pp. 277-290

Larrick, J.W.; Yost, J.A. & Kaplan, J. (1978). Snake bite among the Waorani Indians of Eastern Ecuador. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol.72, No.5, pp. 542-543

León, G.; Rodríguez, M.A.; Rucavado, A.; Lomonte, B. & Gutiérrez, J.M. (2007). Anti-human erythrocyte antibodies in horse-derived antivenoms used in the treatment of snakebite envenomations. *Biologicals*, Vol.35, No.1, pp. 5-11

León, G.; Segura, A.; Herrera, M.; Otero, R.; França, F.O.S.; Barbaro, K.C.; Cardoso, J.L.C.; Wen, F.H.; de Medeiros, C.R.; Prado, J.C.; Málaque, C.M.; Lomonte, B. & Gutiérrez, J.M. (2008). Human heterophylic antibodies against equine immunoglobulins: assessment of their role in the early adverse reactions to antivenom administration.

Transactions of the Royal Society of Tropical Medicine and Hygiene, Vol.102, No.11, pp. 1115-1119

Leynaud, G.C. & Reati, G.J. (2009). Identificación de las zonas de riesgo ofídico en Córdoba, Argentina, mediante el programa SIGEpi. *Revista Panamericana de Salud Pública*, Vol.26, No.1, pp. 64-69

Lomonte, B.; Angulo, Y. & Calderón, L. (2003). An overview of lysine-49 phospholipase A_2 myotoxins from crotalid snake venoms and their structural determinants of myotoxic action. *Toxicon*, Vol.42, No.8, pp. 885-901.

Lomonte, B.; Escolano, J.; Fernández, J.; Sanz, L.; Angulo, Y.; Gutiérrez, J.M. & Calvete, J.J. (2008). Snake venomics and antivenomics of the arboreal neotropical pitvipers *Bothriechis lateralis* and *Bothriechis schlegelii*. *Journal of Proteome Research*, Vol.7, No.6, pp. 2445-2457.

Lomonte, B.; León, G.; Angulo, Y.; Rucavado, A. & Núñez, V. (2009). Neutralization of *Bothrops asper* venom by antibodies, natural products and synthetic drugs: contributions to understanding snakebite envenomings and their treatment. *Toxicon*, Vol.54, No.7, pp. 1012-1028

LoVecchio, F.; Klemens, J.; Roundy, E.B. & Klemens, A. (2003). Serum sickness following administration of Antivenin (Crotalidae) Polyvalent in 181 cases of presumed rattlesnake envenomation. *Wilderness and Environmental Medicine*, Vol. 14, No.4, pp. 220-221

Mackessy, S.P. (2002). Biochemistry and pharmacology of colubrid snake venoms. *Journal of Toxicology-Toxin Reviews*, Vol.21, No. 1-2, pp. 43-83.

Malasit, P.; Warrell, D.A.; Chanthavanich, P.; Viravan, C.; Mongkolsapaya, J.; Singhthong, B. & Supich, C. (1986). Prediction, prevention, and mechanism of early (anaphylactic) antivenom reactions in victims of snake bites. *British Medical Journal*, Vol. 292, pp. 17-20

Manock, S.R.; Suarez, G.; Graham, D.; Avila-Agüero, M.L. & Warrell, D.A. (2008). Neurotoxic envenoming by South American coral snake (*Micrurus lemniscatus helleri*): case report from eastern Ecuador and review. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol.102, No.11, pp. 1127-1132

Meier, J. & Stocker, K.F. (1995). Biology and distribution of venomous snakes of medical importance and the composition of snake venoms, In: *Handbook of Clinical Toxicology of Animal Venoms and Poisons*, J. Meier & J. White, (Eds.), 367-412, CRC Press, ISBN 0-8493-4489-1, Boca Raton, USA

Meyer, W.P.; Habib, A.G.; Onayade, A.A.; Yakubu, A.; Smith, D.C.; Nasidi, A.; Daudu, I.J.; Warrell, D.A. & Theakston, R.D.G. (1997). First clinical experiences with a new ovine Fab *Echis ocellatus* snake bite antivenom in Nigeria: randomized comparative trial with Institute Pasteur Serum (IPSER) Africa Antivenom. *American Journal of Tropical Medicine and Hygiene*, Vol.56, No.3, pp. 291-300

Michael, G.C.; Thacher, T.D. & Shehu, M.I.L. (2010). The effect of pre-hospital care for venomous snake bite on outcome in Nigeria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol.105, No.2, pp. 95-101

Mohapatra, B.; Warrell, D.A.; Suraweera, W.; Bhatia, P.; Dhingra, N.; Jotkar, R.M.; Rodriguez, P.S.; Mishra, K.; Whitaker, R. & Jha, P. (2011). Snakebite mortality in India: a nationally representative mortality survey. *PLoS Neglected Tropical Diseases*, Vol.5, No.4, pp. e1018

Moura-da-Silva, A.M.; Serrano, S.M.T.; Fox, J.W. & Gutiérrez, J.M. (2009) Snake venom metalloproteinases. Structure, function and effects on snake bite pathology. In: *Animal Toxins: State of the Art. Perspectives in Health and Biotechnology*, M.E. de Lima;

A.M.C. Pimenta; M.F. Martin-Euclaire; R.B. Zingali & H. Rochat, (Eds.), 525-546, Editora UFMG, ISBN 978-85-7041-735-0, Belo Horizonte, Brazil.

Ohno, M.; Chijiwa, T.; Oda-Ueda, N.; Ogawa, T. & Hattori, S. (2003). Molecular evolution of myotoxic phospholipases A_2 from snake venom. *Toxicon*, Vol.42, No.8, pp. 841-854

Otero, R.; Valderrama, R.; Osorio, R.G. & Posada, L.E. (1992). Programa de atención primaria del accidente ofídico. Una propuesta para Colombia. *Iatreia*, Vol.5, No.2, pp. 96-102

Otero, R.; Núñez, V.; Osorio, R.G.; Gutiérrez, J.M.; Giraldo, C.A. & Posada, L.E. (1995). Ability of six Latin American antivenoms to neutralize the venom of mapaná equis (*Bothrops atrox*) from Antioquia and Chocó (Colombia). Toxicon, Vol.33, No.6, pp. 809-815

Otero, R.; Gutiérrez, J.M.; Rojas, G.; Núñez, V.; Díaz, A.; Miranda, E.; Urige, A.F.; Silva, J.F.; Ospina, J.G.; Medina, Y.; Toro, M.F.; García, M.E.; León, G.; García, M.; Lizano, S.; de la Torre, J.; Márquez, J.; Mena, Y.; González, N.; Arenas, L.C.; Puzón, A.; Blanco, N.; Sierra, A.; Espinal, M.E.; Arboleda, M.; Jiménez, J.C.; Ramírez, P.; Díaz, M.; Guzmán, M.C.; Barros, J.; Henao, S.; Ramírez, A.; Macea, U. & Lozano, R. (1999). A randomized blinded clinical trial of two antivenoms, prepared by caprylic acid or ammonium sulphate fractionation of IgG, in *Bothrops* and *Porthidium* snake bites in Colombia: correlation between safety and biochemical characteristics of antivenoms. *Toxicon*, Vol.37, No.6, pp. 895-908

Otero, R.; Fonnegra, R.; Jiménez, S.L.; Núñez, V.; Evans, N.; Alzate, S.P.; García, M.E.; Saldarriaga, M.; del Valle, G.; Osorio, R.G.; Díaz, A.; Valderrama, R.; Duque, A. & Vélez, H.N. (2000). Snakebites and ethnobotany in the northwestern region of Colombia: Part I: traditional use of plants. *Journal of Ethnopharmacology*, Vol.71, No.3, pp. 493-504

Otero, R.; Gutiérrez, J.; Mesa, M.B.; Duque, E.; Rodríguez, O.; Arango, J.L.; Gómez, F.; Toro, A.; Cano, F.; Rodríguez, L.M.; Caro, E.; Martínez, J.; Cornejo, W.; Gómez, L.M.; Uribe, F.L.; Cárdenas, S.; Núñez, V. & Díaz, A. (2002). Complications of *Bothrops*, *Porthidium*, and *Bothriechis* snakebites in Colombia. A clinical and epidemiological study of 39 cases attended in a university hospital. *Toxicon*, Vol.40, No.8, pp. 1107-1114

Otero-Patiño, R.; Cardoso, J.L.C.; Higashi, H.G.; Núñez, V.; Díaz, A.; Toro, M.F.; García, M.E.; Sierra, A.; García, L.F.; Moreno, A.M.; Medina, M.C.; Castañeda, N.; Silva-Díaz, J.F.; Murcia, M.; Cárdenas, S.Y. & Dias-da-Silva, W. (1998). A randomized, blinded, comparative trial of one pepsin-digested and two whole IgG antivenoms for *Bothrops* snake bites in Urabá, Colombia. *American Journal of Tropical Medicine and Hygiene*, Vol. 58, No.2, pp. 183-189

Perales, J.; Neves-Ferreira, A.G.; Valente, R.H. & Domont, G.B. (2005). Natural inhibitors of snake venom hemorrhagic metalloproteinases. *Toxicon*, Vol.45, No.8, pp. 1013-1020

Petras, D.; Sanz, L.; Segura, A.; Herrera, M.; Villalta, M.; Solano, D.; Vargas, M.; León, G.; Warrell, D.A.; Theakston, R.D.G.; Harrison, R.A.; Durfa, N.; Nasidi, A.; Gutiérrez, J.M. & Calvete, J.J. (2011). Snake venomics of African spitting cobras: toxin composition and assessment of congeneric cross-reactivity of the pan-African EchiTAb-Plus-ICP antivenom by antivenomics and neutralization approaches. J*ournal of Proteome Research*, Vol.10, No.3, pp. 1266-1280

Pierini, S.V.; Warrell, D.A.; de Paulo, A. & Theakston, R.G.D. (1996). High incidence of bites and stings by snakes and other animals among rubber tappers and Amazonian Indians of the Juruá Valley, Acre State, Brazil. *Toxicon*, Vol.34, No.2, pp. 225-236

Prasarnpun, S.; Walsh, J.; Awad, S.S. & Harris, J.B. (2005). Envenoming bites by kraits: the biological basis of treatment-resistant neuromuscular paralysis. *Brain*, Vol. 128, No.12, pp. 2987-2996

Pugh, R.N. & Theakston, R.D.G. (1980). Incidence and mortality on snake bite in savanna Nigeria. *Lancet*, Vol.2, pp. 1181-1183

Pugh, R.N.; Theakston, R.D.G. & Reid, H.A. (1980). Malumfashi Endemic Diseases Research Project, XIII. Epidemiology of human encounters with the spitting cobra, *Naja nigricollis*, in the Malumfashi area of northern Nigeria. *Annals of Tropical Medicine and Parasitology*, Vol.74, No.5, pp. 523-530

Quijada-Mascareñas, A. & Wüster, W. (2010). Recent advances in venomous snake systematics, In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 25-64, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Rahman, R.; Faiz, M.A.; Selim, S.; Rahman, B.; Basher, A.; Jones, A.; d'Este, C.; Hossain, M.; Islam, Z.; Ahmed, H. & Milton, A.H. (2010). Annual incidence of snake bite in rural Bangladesh. *PLoS Neglected Tropical Diseases*, Vol.4, No.10, pp. e860

Raw, I.; Guidolin, R.; Higashi, H.G & Kelen, E.M.A. (1991). Antivenins in Brazil: Preparation. In: *Handbook of Natural Toxins, Vol 5, Reptile Venoms and Toxins,* A.T. Tu (Ed,), 557-581, Marcel Dekker, New York, USA.

Rodrigues-Silva, R.; Antunes, G.F.; Velarde, D.T. & Santoro, M.M. (1999). Thermal stability studies of hyperimmune horse antivenoms. *Toxicon*, Vol.37, No.1, pp. 33-45

Rojas, G.; Jiménez, J.M & Gutiérrez, J.M. (1994). Caprylic acid fractionation of hyperimmune horse plasma: description of a simple procedure for antivenom production. *Toxicon*, Vol.32, No.3, pp. 351-363

Rossetto, O.; Morbiato, L.; Caccin, P.; Rigoni, M. & Montecucco, C. (2006). Presynaptic enzymatic neurotoxins. *Journal of Neurochemistry*, Vol.97, No.6, pp. 1534-1545

Rucavado, A.; Escalante, T.; Franceschi, A.; Chaves, F.; León, G.; Cury, Y.; Ovadia, M. & Gutiérrez, J.M. (2000). Inhibition of local hemorrhage and dermonecrosis induced by *Bothrops asper* snake venom: effectiveness of early *in situ* administration of the peptidomimetic metalloproteinase inhibitor batimastat and the chelating agent $CaNa_2EDTA$. *American Journal of Tropical Medicine and Hygiene*, Vol.63, No.5-6, pp. 313-319.

Saravia, P.; Rojas, E.; Arce, V.; Guevara, C.; López, J.C.; Chaves, E.; Velásquez, R.; Rojas, G. & Gutiérrez, J.M. (2002). Geographic and ontogenetic variability in the venom of the neotropical rattlesnake *Crotalus durissus*: pathophysiological and therapeutic implications. *Revista de Biología Tropical*, Vol. 50, No.1, pp. 337-346

Saul, M.E.; Thomas, P.A.; Dosen, P.J.; Isbister, G.K.; O'Leary, M.A.; Whyte, I.M.; McFadden, S.A. & van Heyden, D.F. (2011). A pharmacological approach to first aid treatment for snakebite. *Nature Medicine*, Vol.17, No.7, pp. 809-811

Scherrmann, J.M. (1994). Antibody treatment of toxin poisoning-recent advances. *Journal of Toxicology-Clinical Toxicology*, Vol.32, No.4, pp. 363-375

Segura, A.; Herrera, M.; González, E.; Vargas, M.; Solano, G.; Gutiérrez, J.M & León, G. (2009). Stability of equine IgG antivenoms obtained by caprylic acid precipitation: towards a liquid formulation stable at tropical room temperature. *Toxicon*, Vol.53, No.6, pp. 609-615

Segura, A.; Castillo, M.C.; Núñez, V.; Yarlequé, A.; Gonçalves, L.R.; Villalta, M.; Bonilla, C.; Herrera, M.; Vargas, M.; Fernández, M.; Yano, M.Y.; Araújo, H.P.; Boller, M.A.; León, P.; Tintaya, B.; Sano-Martins, I.S.; Gómez, A.; Fernández, G.P.; Geoghegan, P.; Higashi, H.G., León, G. & Gutiérrez, J.M. (2010a). Preclinical assessment of the neutralizing capacity of antivenoms produced in six Latin American countries

against medically-relevant *Bothrops* snake venoms. *Toxicon*, Vol.56, No.6, pp. 980-989

Segura, A.; Villalta, M.; Herrera, M.; León, G.; Harrison, R.; Durfa, N.; Nasidi, A.; Calvete, J.J.; Theakston, R.D.G.; Warrell, D.A. & Gutiérrez, J.M. (2010b). Preclinical assessment of the efficacy of a new antivenom (EchiTAb-Plus-ICP) for the treatment of viper envenoming in sub-Saharan Africa. *Toxicon*, Vol.55, No.2-3, pp. 369-374

Serrano, S.M.T. & Maroun, R.C. (2005). Snake venom serine proteinases: sequence homology vs. substrate specificity, a paradox to be solved. *Toxicon*, Vol.45, No.8, pp. 1115-1132.

Sharma, S.K.; Chappuis, F.; Jha, N.; Bovier, P.A.; Loutan, L. & Koirala, S. (2004). Impact of snake bites and determinants of fatal outcomes in southeastern Nepal. *American Journal of Tropical Medicine and Hygiene*, Vol.21, No.2, pp. 234-238

Simpson, I.D. & Norris, R.L. (2007). Snake antivenom product guidelines in India: "the devil is in the details". *Wilderness and Environmental Medicine*, Vol.18, No.3, pp. 163-168

Simpson, I.D. (2008). A study of the current knowledge base in treating snake bite amongst doctors in the high-risk countries of India and Pakistan: does snake bite treatment training reflect local requirements? *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol. 102, No.11, pp. 1108-1114

Smalligan, R.; Cole, J.; Brito, N.; Laing, G.D.; Mertz, B.L.; Manock, S.; Maudin, J.; Quist, B.; Holland, G.; Nelson, S.; Lalloo, D.G.; Rivadeneira, G.; Barragan, M.E.; Dolley, D.; Addleston, M.; Warrell, D.A. & Theakston, R.D.G. (2004). Crotaline snake bite in the Ecuadorian Amazon: randomised double blind comparative trial of three South American polyspecific antivenoms. *British Medical Journal*, Vol.328, pp. 1129

Snow, R.W.; Bronzan, R.; Roques, T.; Nyamawi, C.; Murphy, S. & Marsh, K. (1994). The prevalence and morbidity of snake bite and treatment-seeking behavior among a rural Kenyan population. *Annals of Tropical Medicine and Parasitology*, Vol.88, No.6, pp. 665-671

St Pierre, L.; Masci, P.P.; Filippovich, I.; Sorokina, N.; Marsh, N.; Miller, D.J. & Lavin, M.F. (2005). Comparative analysis of prothrombin activators from the venom of Australian elapids. *Molecular Biology and Evolution*, Vol.22, No.9, pp. 1853-1864

Sutherland, S.K. (1977). Serum reactions. An analysis of commercial antivenoms and the possible role of anticomplementary activity in de-novo reactions to antivenoms and antitoxins. *Medical Journal of Australia*, Vol.1,No.17, pp. 613-615

Sutherland, S.K.; Coulter, A.R. & Harris, R.D. (1979). Rationalisation of first-aid measures for elapid snakebite. *Lancet*, Vol.1, pp. 183-186

Swaroop, S. & Grab, B. (1954). Snakebite mortality in the world. *Bulletin of the World Health Organization*, Vol.10, No.1, pp. 35-76

Tanaka, G.D.; Furtado, M.F.; Portaro, F.C.; Sant'Anna, O.A. & Tambourgi, D.V. (2010). Diversity of *Micrurus* snake species related to their venom toxic effects and the prospective of antivenom neutralization. *PLoS Neglected Tropical Diseases*, Vol.4, No.3, pp. e622

Tans, G. & Rosing, J. (2001). Snake venom activators of factor X: an overview. *Haemostasis*, Vol.31, No.3-6, pp. 225-233

Theakston, R.D.G. (1986). Characterization of venoms and standardization of antivenoms. In: *Natural Toxins. Animal, Plant and Microbial*, J.B. Harris, (Ed.), 287-303, Clarendon Press, Okford, United Kingdom.

Theakston, R.D.G.; Warrell, D.A. & Griffiths, E. (2003). Report of a WHO workshop on the standardization and control of antivenoms. *Toxicon*, Vol.41, No.5, pp. 541-557

Thomas, L.; Tyburn, B. & the Research Group on Snake Bite in Martinique (1996). *Bothrops lanceolatus* bites in Martinique: Clinical aspects and treatment. In: *Envenomings and Their Treatments*, C. Bon & M. Goyffon, (Eds), 255-265, Fondation Marcel Mérieux, Lyon, France.

Trape, J.F.; Pison, G.; Guyavarch, E. & Mane, Y. (2001). High mortality from snakebite in south-eastern Senegal. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol.95, No.4, pp. 420-423

Trinh, K.X.; Khac, Q.L.; Trinh, L.X. & Warrell, D.A. (2010). Hyponatremia, rhabdomyolysis, alterations in blood pressure and persistent mydriasis in patients envenomed by Malayan kraits (*Bungarus candidus*) in southern Viet Nam. *Toxicon*, Vol.56, No.6, pp. 1070-1075

Tun-Pe; Phillips, R.E.; Warrell, D.A.; Moore, R.A.; Tin-Un-Swe; Myint-Lwin & Burke, C.W. (1987). Acute and chronic pituitary failure resembling Shehan's syndrome following bites by Russell's viper in Burma. *Lancet*, Vol.2, pp. 763-767

Visser, L.E.; Kyed-Faried, S.; Belcher, D.W.; Geelhoed, D.W.; van Leeuwen, J.S. & van Roosmalen, J. (2008). Failure of a new antivenom to treat *Echis ocellatus* snake bite in rural Ghana: the importance of quality surveillance. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, Vol. 102, No.5, pp. 445-450

Vonk, F.J.; Admiraal, J.F.; Jackson, K.; Reshef, R.; de Bakker, M.A.; Vanderschoot, K.; vanden Berge, I.; van Atten, M.; Burgerhout, E.; Beck, A.; Mirtschin, P.J.; Kochva, E.; Witte, F.; Fry, B.G.; Woods, A.E. & Richardson, M.K. (2008). Evolutionary origin and development of snake fangs. *Nature*, Vol.454, pp. 630-633

Warrell, D.A.; Davidson, N.M.; Omerod, L.D.; Pope, H.M.; Watkins, B.J.; Greenwood, B.M. & Ried, H.A. (1974). Bites by the saw-scaled or carpet viper (*Echis carinatus*): trial of two specific antivenoms. *British Medical Journal*, Vol.4, pp. 437-440

Warrell, D.A. (1992) The global problem of snake bite: its prevention and treatment. In: *Advances in Toxinology Research, Vol. 1,* P. Gopalakrishnakone & C.K. Tan, (Eds), 121-153, National University of Singapore, Singapore.

Warrell, D.A. (1995a). Clinical toxicology of snakebite in Asia, In: *Handbook of Clinical Toxicology of Animal Venoms and Poisons*, J. Meier & J. White, (Eds.), 493-594, CRC Press, ISBN 0-8493-4489-1, Boca Raton, USA

Warrell, D.A. (1995b). Clinical toxicology of snakebite in Africa and the Middle East / Arabian Peninsula, In: *Handbook of Clinical Toxicology of Animal Venoms and Poisons*, J. Meier & J. White, (Eds.), 433-492, CRC Press, ISBN 0-8493-4489-1, Boca Raton, USA

Warrell, D.A. (1996). Clinical features of envenoming from snake bites. In: *Envenomings and Their Treatments*, C. Bon & M. Goyffon, (Eds), 63-76, Fondation Marcel Mérieux, Lyon, France.

Warrell, D.A. (1997). Geographical and intraspecies variation in the clinical manifestationsof envenoming by snakes. In: *Venomous Snakes. Ecology, Evolution and Snakebite*, R.S.Thorpe, W. Wüster & A. Malhotra, (Eds.), 189-203, Clarendon Press, Oxford, United Kingdom.

Warrell, D.A. (2004). Snakebites in Central and South America: epidemiology, clinical features and clinical management, In: *The Venomous Reptiles of the Western Hemisphere*, J.A. Campbell & W.W. Lamar, (Eds.), 709-761, Cornell University Press, ISBN 0-8014-4141-2, Ithaca, USA

Warrell, D.A. (2010). Snake bite. *Lancet*, Vol.375, pp. 77-88

White, J. (2010). Envenomation. Prevention and treatment in Australia. In: *Handbook of Venoms and Toxins of Reptiles*, S.P. Mackessy, (Ed.), 423-451, CRC Press, ISBN 978-0-8493-9165-1, Boca Raton, USA

Williams, D.; Gutiérrez, J.M.; Harrison, R.A.; Warrell, D.A., White, J.; Winkel, K.D. & Gopalakrishnakone, P. (2010). The Global Snake Bite Initiative: an antidote for snake bite. *Lancet,* Vol.375, pp. 89-91

Williams, D.J.; Gutiérrez, J.M.; Calvete, J.J.; Wüster, W.; Ratanabanangkoon, K.; Paiva, O.; Brown, N.I.; Casewell, N.R.; Harrison, R.A.; Rowley, P.D.; O'Shea, M., Jensen, S.D.; Winkel, K.D. & Warrell, D.A. (2011). Ending the drought: new strategies for improving the flow of affordable, effective antivenoms in Asia and Africa. *Journal of Proteomics*, Vol.74, No.9, pp. 1735-1767

Williams, S.S.; Wijesinghe, C.A.; Jayamanne, S.F.; Buckley, N.A.; Dawson, A.H.; Lalloo, D.G. & de Silva, H.J. (2011). Delayed psychological morbidity associated with snakebite envenoming. *PLoS Neglected Tropical Diseases*, Vol.5, No.8, pp. e1255

World Health Organization (2007a). *Rabies and Envenomings. A Neglected Public Health Issue*, World Health Organization, ISBN 978 92 4 156348 2, Geneva, Switzerland

World Health Organization (2007b). *International Statistical Classification of Diseases and Related Health Problems*, *10th Revision*, World Health Organization, Geneva, Switzerland, Available from http://apps.who.int/classifications/apps/icd/icd10online/

World Health Organization (2007c). *Global Plan to Combat Neglected Tropical Diseases 2008-2015*, World Health Organization, Geneva, Switzerland, Available from http://whqlibdoc.who.int/hq/2007/WHO_CDS_NTD_2007.3_eng.pdf

World Health Organization (2010a). *WHO Guidelines for the Production, Control and Regulation of Snake Antivenom Immunoglobulins,* World Health Organization, Geneva, Switzerland, Available from http://www.who.int/bloodproducts/snake_antivenoms/snakeantivenomguide/en/

World Health Organization (2010b). *Guidelines for the Prevention and Clinical Management of Snakebite in Africa*, World Health Organization, Geneva, Switzerland, Available from http://www.afro.who.int/en/clusters-a-programmes/hss/essential-medicines/highlights/2731-guidelines-for-the-prevention-and-clinical-management-of-snakebite-in-africa.html

World Health Organization (2011). *WHO Model List of Essential Medicines*, World Health Organization, Geneva, Switzerland. Available from http://www.who.int/medicines/publications/essentialmedicines/en/index.html

Yingprasertchai, S.; Bunyasrisawat, S. & Ratanabanangkoon, K. (2003). Hyaluronidase inhibitors (sodium chromoglycate and sodium auro-thiomalate) reduce the local tissue damage and prolong the survival time of mice injected with *Naja kaouthia* and *Calloselasma rhodostoma* venoms. *Toxicon*, Vol.42, No.6, pp. 635-646

8

Chemical Residues in Animal Food Products: An Issue of Public Health

María Constanza Lozano and Mary Trujillo
Pharmacy Department, Faculty of Sciences, National University of Colombia
Colombia

1. Introduction

Human beings consume protein-rich foods to supply their nutritional requirements, mainly of animal origin, such origin lying in meat from different species (cattle, sheep, caprines, birds, pigs, fish and seafood/shellfish), milk and eggs. With the exception of some products derived from fishing, these foods are obtained from financial exploitations in which the animals' health must be guaranteed, thereby ensuring that food is harmless. In several countries the safety of such food has mainly been focused on avoiding the transmission of zoonotic diseases, less attention thus being paid to potentially present chemical residues, perhaps due to the course of the resulting disease. Whilst infectious processes are frequently of the acute type, toxicosis caused by contaminants in foods (more than acute) may be chronic, silent and often lacking a known aetiological agent.

The primary production of such food involves the animals' interaction with their setting from which they may become exposed to undesirable chemical substances which may generate residuality. The chemical substances to which animals may become exposed during their production cycle which have been identified to date could come from drugs and growth promoters aimed at treating diseases and improving production parameters, biologically-derived toxins (mycotoxins, phycotoxins, phytotoxins) and/or environmental contaminants linked to atmospheric pollution, from the soil and/or water. This chapter will be orientated towards dealing with residues from chemicals substances in foodstuffs of animal origin caused by drugs and growth promoters, as well as by toxins having a biological origin. It will also deal with general concepts such as toxic agent's target in an organism, the regulation of residues in food and the analytical methods used for detecting them. The contamination of food by chemical risks is a worldwide public health matter which may also hamper international trade.

2. The destination for toxic agents in an organism

Living beings continuously are being exposed to external substances, generically called xenobiotics, which can have adverse effects according to their chemical characteristics. Oral, dermal and inhalation routes represent the commonest means of exposure to these substances, the first being of interest as it deals with risks to human health due to the consumption of foodstuffs contaminated by potentially toxic substances. On the other hand, animals (representing a readily available source of food for humans) are exposed to

xenobiotics in multiple ways which could be present in their products. If one is dealing with veterinary drugs then the route of exposure could be oral (for example, coccidiostatics in poultry), dermal (e.g. external antiparasitic agents in ruminants), parenteral (e.g. antibiotic treatment in large animals) and even inhalation (if the animals are given anaesthesia before surgical procedures). Biologically-derived toxins mainly enter food-producing animals by the oral route (e.g. forage contaminated with mycotoxins or fish consuming toxic algae).

Xenobiotics in an organism go through a series of stages including absorption, distribution, metabolism and excretion, forming part of the pharmacokinetics or toxicokinetics according to the effects produced by a particular substance (pharmacological or toxicological). Xenobiotics enter a food-producing animals' organism and, according to its kinetics, reach the tissues which will become food for human beings (Figure 1). These concepts will be dealt with below, approaching them from the perspective of potential residuality which different substances can cause in an animal's organism.

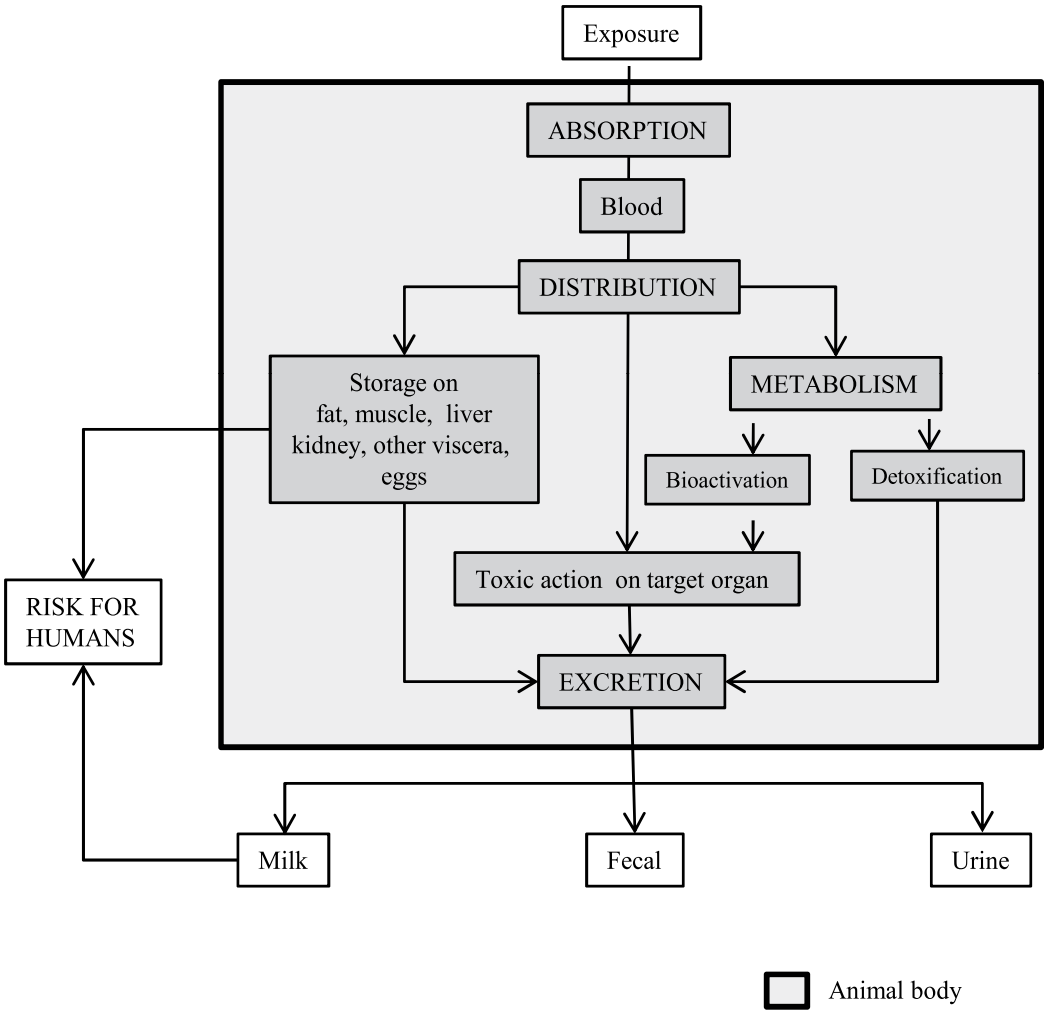


Fig. 1. Destination of toxicants in foodstuffs animal organism and risk for humans

2.1 Absorption

During absorption, the xenobiotics cross cell membranes and reach the systemic circulation. Many toxic agents enter with foodstuff and are absorbed by the same routes as the other substances present in them. The chemicals cross the cell membrane's lipid bilayer through two basic processes: diffusion (favouring their concentration gradients) and active transport (against their concentration gradients). Most liposoluble xenobiotics are transported by simple diffusion through the cell membranes. Organic acids and/or bases thus tend to become absorbed when they are in their most liposoluble form (non-ionised), which is determined by surrounding pH. Weak acids will become more easily absorbed in the stomach whilst this will happen to bases in the intestines. Hydrophilic substances having a small molecular weight become diffused through aqueous pores formed by proteins (facilitated diffusion). Active transport, against concentration gradients requiring an expenditure of energy, occurs through proteins present on the membrane mobilising a substance from one side to another (Lehman-McKeeman, 2008).

It should be born in mind that certain factors may sometimes alter xenobiotics' absorption; for example, the flora present in the gastrointestinal tract may transform them and leave them less available for being absorbed. This is why ruminants are resistant to some mycotoxins. Pre-systemic elimination of a toxic agent may occur with enterohepatic circulation, thereby minimising its potential adverse effect.

2.2 Distribution

Once it has been absorbed, a toxic agent becomes distributed throughout the whole body; during its initial phase this distribution is dominated by the blood flow. The penetration of toxic agent into the cells depends on passive diffusion or specialised transport; however, certain toxicants do not cross the membranes and become distributed via the blood flow. Some become accumulated in determined parts of the organism as a result of their binding to proteins or their high solubility in fat. When a toxicant becomes stored, then equilibrium is reached with the free fraction which is in the plasma. Thus, when the chemical becomes metabolised or is excreted, then substance is released from the storage site, thereby meaning that the xenobiotic half-life could become very long.

Albumin is the main plasmatic protein transporting xenobiotics. This protein may also be a toxicant reservoir since it impedes transport through the membranes due to its high molecular weight. The presence of toxic agents in the blood could be exploited for recognising exposure, whether in humans or animals.

Many organic compounds are very stable and lipophilic, becoming accumulated in the environment, becoming rapidly absorbed and concentrated in body fat. The toxicants become accumulated in fat because they are dissolved in it. A substance stored in fat is not toxic for the carrier, but there is rapid lipid mobilisation, for example poisoning may occur during long periods of fasting. Animal fat, a potential reserve of liposoluble toxicants, could be consumed by human beings.

The liver and the kidneys have a large capacity for proteins to bind a broad range of chemicals. These organs important function lying in the metabolism and elimination of xenobiotics makes them concentrate more toxic agents than all the rest combined. Thus consuming such viscera may represent a risk for the end consumer. There is a lower

presence of residues in animals' musculature (meat) compared to the viscera (kidneys, liver) and fat. Their accumulation at an injection site is feasible in cases where there has been exposure to the intramuscular drug route, this being important in animals which are to be consumed by humans.

The distribution of some chemicals in eggs, as well as reducing palatability, could represent a risk for the end consumer.

2.3 Metabolism

The object of xenobiotics' metabolism is to increase characteristics regarding an increase in substances' hydrosolubility so that they can be more easily excreted. This process occurs in two phases; hydrolysis, reduction and oxidation reactions are presented during phase 1, most of them being enzyme-mediated. Cytochrome P450 (CYP450) oxidation enzymes being of particular importance during this phase due to their catalytic versatility and the great number of xenobiotics constituted in their substrate. Conjugation reactions occur during phase 2, mainly with glucuronic acid, glutathione conjugates and sulphates; such reactions are enzymatically mediated by protein superfamilies called, respectively, uridine diphosphate glucuronosyltransferase, glutathione S-transferases and sulphotransferases. In spite of the initial purpose of xenobiotics' metabolism (or biotransformation) being detoxification, substances can occasionally acquire their true toxic power on being biotransformed; such reaction is called bioactivation or metabolic activation. Aflatoxins and pyrrolizidine alkaloids are bioactivated substances of interest regarding the residuality which they represent in food of animal origin.

2.4 Excretion

Toxicants are eliminated from the body by various routes, the kidneys being the most important organ for excreting xenobiotics since it is the main elimination route. The biliary route involving the faeces is the other elimination route for toxic substances which have been consumed.

Milk is an important elimination route due to the risk of contamination which it represents; this liquid is a lipid emulsion in an aqueous protein solution and may thus contain whatever toxicant which is in solution in an animal's body water. Simple chemicals arrive at the mammary glands by diffusion in their free form, bound to proteins or dissolved in lipids. The percentage of the total amount of compounds eliminated in milk is very low because the other elimination routes are more efficient. However, the main problem lies in chronic exposure and/or liposoluble compounds (Panter & James, 1990).

The concept of withdrawal time has been established to avoid the accumulation of drug residues in animals; it is defined as being the time required after a drug has been administered to an animal to ensure that drug residues in marketable products (meat, eggs, viscera or other edible products) are below a determined maximum residue limit (MRL).

3. Regulating and evaluating risk

There can never be an absolute guarantee that our food is safe; it is simply impossible to test every contaminant. Every country has an agency which oversees food safety; this is defined

as being the, "reasonable certainty of no harm," and the aforementioned agencies regulate which additives are allowed in food and what levels of unavoidable contaminants are acceptable. The US Department of Agriculture's (USDA) Food Safety Inspection Service (FSIS) is responsible for the safety of meat, poultry, and egg products in the USA (Lodovico et al., 2008). The European Food Safety Authority (EFSA) is the keystone of the European Union's (EU) risk assessment regarding food and animal feed safety. The Codex Alimentarius Commission (created by the FAO and WHO) develops food standards, guidelines and related texts such as codes of practice under the Joint FAO/WHO Food Standards Programme (JECFA). The main purposes of this programme are protecting consumers' health, ensuring fair trade practices in the food trade and promoting the coordination of all food standards' work undertaken by international governmental and non-governmental organisations.

Health authorities recommend maximum acceptable or tolerable levels for chemicals which are neither genotoxic nor carcinogenic, such as acceptable daily intake (ADI), reference dose (RfD), especially for pesticides, tolerable daily intake (TDI) and provisional tolerable weekly intake (PTWI) for contaminants which may accumulate in the body. The responsible agencies conduct risk assessment to determine such levels; this consists of hazard identification and characterisation, exposure assessment and subsequent risk characterisation.

Hazards are identified and characterised from human epidemiological observations and animal-based toxicity testing supported by *in vitro* mechanistic studies which can make extrapolation from animals to humans become more realistic. Structure–activity relationships-based indications and the increased use of novel molecular biology techniques are also very valuable.

Dose–response information is essential for quantifying an adverse health effect. This may be graphically presented as being the relationship between the increase of a dose and the increase of a pertinent biological response. Such dose–response curve is essential for identifying a non-active dose taken as being the no observed adverse effect level (NOAEL), the highest dose of a substance which causes no detectable adverse alteration in line with defined treatment conditions. Interspecies differences should be taken into account as well as the fact that humans may exhibit substantial differences in their sensitivity to certain toxins due to differences regarding metabolic pathways and other factors. Uncertainty factors are thus applied when extrapolating from the toxicity observed in laboratory animals to health risks in humans, this usually being a factor of 10 for interspecies difference and a factor of up to 10 for human variability (depending on the extent and quality of available human data).

The resulting value (equation 1) provides an estimate of the amount of a substance in food, expressed on a body weight basis, that can be ingested daily over a lifetime without appreciable risk (standard human= 60 kg). The ADI is then used for determining the maximum allowable levels of a particular chemical in a specific food, depending on the extent to which this food contributes towards the overall intake of such chemical (Lodovico et al., 2008). These are called maximum limits for some chemicals and maximum residue limits (MRLs) for substances such as veterinary drugs and hormone residues.

$$ADI = \frac{NOAEL}{UF}$$

ADI : Acceptable Daily Intake
NOAEL : Not Observed Adverse Effect Level
UF : Uncertainty Factor (10, 100, n)

Equation 1. Acceptable daily intake calculation

The regulatory approach (not strictly scientific) for genotoxic and carcinogenic compounds is based on the assumption that there is no threshold dose (it is assumed that one genotoxic molecule is sufficient to hit a single DNA base, thereby inducing damage). The aim in this case is to keep the exposure level as low as technologically achievable (Lodovico et al., 2008).

4. Analytical methodologies

Analytical data quality is a key factor in the success of a control programme dealing with residues in foodstuffs. The analytical results of methods regarding official standards offer the necessary information for developing and managing programmes responding to a population's public health needs. It is very important that sanitary authorities have readily available practical analysis methods which will reliably detect and quantify (without ambiguity) a drug's residues which could be present in meat, milk, or eggs at a suitable concentration level. Unfortunately, methods having these attributes are not available for all residues, partly due to the large amount of possible substances which may be found in animals' food chains.

Chemical residues in food of animal origin, such as meat, milk or eggs, are frequently present in very low concentrations or trace levels, thereby representing an important challenge for a chemical analyst, given that the analytical methods developed must be highly sensitive and selective.

The prior treatment which a sample has received is very important for ensuring that methods reach desired detection levels as well as an acceptable level of exactitude and precision, thereby enabling the factors responsible for analyte loss to be controlled during such procedure (The Spanish Industrial Pharmaceutics Association – Asociación Española de Farmacéuticos de la Industria [AEFI], 2001); this would include the presence of functional reactive groups which can interfere with such determination (LoBrutto & Patel, 2007).

Analytical methods will always have to be tested/proved on material from each animal species since differences in composition (fat, specifics proteins, eg: myoglobin) can influence both analyte extraction and separation. Another important consideration concerns the treatment which biotransformation enzyme-rich tissues such as the liver should receive as this may induce post-mortem metabolism thereby altering real results. Special management musts also be used for determining analytes in eggs because eggs consist of two distinct compartments (the white and the yolk) whose chemical composition is different, as well as depending on the components of chicken's diet.

The scientific community focuses on developing reliable, economic and rapid methods (and whenever possible automated) which could be applied to evaluating the safety of foodstuffs bearing in mind the broad range of existing chemicals and matrices (Botsoglou & Fletouris, 2000). A unified procedure would eliminate the need for using separate methods for detecting multiple residues in the same product; however, such methods are rarely found in real life. There is an immense variety of methods for identifying, confirming and quantifying analytes which could be used individually or coupled to each other in a suitable way. These methods can be grouped into bioassays, microbiology assays, immunochemical assays and physical-chemical assays.

4.1 Bioassays

Biological methods for determining toxic residues in foodstuffs can be used both *in vivo* and *in vitro* (FAO, 2004) and have been particularly developed for detecting and quantifying the phycotoxins present in shellfish. The mouse bioassay is the most used one and is even accepted by regulating entities.

A toxin extract is intraperitoneally injected into mice having around 20 g body weight in the mouse bioassay and their survival is monitored from 24 to 48 hours. One mouse unit (MU) is defined as being the minimum quantity of toxin needed to kill a mouse within 24 hours. Sample toxicity (MU/g whole tissue) is determined from the smallest dose at which two mice or more in a group of three die within 24 hours. The regulatory level is set at 0.05 MU/g whole tissues in many countries; this assay's major disadvantages are therefore a lack of specificity (no differentiation between various toxin components), subjectivity regarding the animals' time of death as well as maintaining and killing laboratory animals. This assay may also give false positives because of interference which can be very toxic for mice. The EU has issued directions on how to perform this assay in an attempt to standardise the mouse bioassay methodology.

Other bioassays which are also used would include the suckling mouse assay for detecting marine toxins which determin the weight of the intestine regarding body weight, the rat bioassay which is based on inducing diarrhoea in rats, the *Daphnia magna* assay which is used for detecting okadaic acid, the intestinal loop assay which determines the accumulation of fluids in rabbit intestine and mice and cytotoxicity assays which are based on observing morphological changes in cells.

4.2 Microbiological assays

The microbiological methods used for detecting antimicrobial residues in foodstuffs are based on inhibiting microbial growth, microbial receptor activity and enzymatic reactions and could be applied to all types of matrices, usually milk, meat, eggs and honey. Microbial inhibition assays involve culturing a microorganism from a standard strain, usually *Bacillus stearothermophilus, Bacillus subtilis, Bacillus cereus, Micrococcus luteus, Escherichia coli, Bacillus megatherium, Sarcina lutea* and/or *Streptococcus thermophilus* (AEFI, 2001).

4.3 Immunochemical assays

Immunochemical methods represent an important tool for determining drug residues, given their high specificity, they lead to analytes being determined in samples having had very

reduced prior cleaning treatment. These assays are based on the reaction of an antigen binding to a specific primary antibody or for each antigen, analogously to an enzyme-substrate reaction. The most common immunochemical methods would include the enzyme-linked immunosorbent assay (ELISA), direct and indirect competitive enzyme-linked immunosorbent assays, immunoaffinity chromatography (IAC), radioimmunoassay (RIA), the enzyme-monitored immunotest (EMIT), the fluorescent immunoassay (FIA) and the chemiluminescence immunoassay.

4.4 Physical-chemical assays

Physical-chemical methods are mainly used for isolating, separating, quantifying and confirming the presence of dangerous residues in samples; this requires that the sensitivity of a particular selection method and the determinative or confirmation method are similar. Numerous procedures based on the analytes'different physicochemical properties have been developed for achieving this objective. Even though a drug's chemical structure greatly determines the most suitable method for its determination, different methods are usually available for the same analyte due to the large amount of possibilities and by coupling different methods to obtain optimum analyte separation and detection.

Separation methods are based on the principles of chromatography and are generally coupled to high sensitivity and selectivity detection techniques leading to quantifying an analyte with a high level of precision and exactitude and also its unequivocal identification at very low concentration levels. The chromatographic methods used for determining analytes in complex matrices would be gas chromatography (GC), high performance liquid chromatography (HPLC), ionic chromatography (IC), size exclusion chromatography (SEC), supercritical fluid chromatography (SFC), affinity chromatography (AC).

Spectrometric methods are also used either alone or coupled to chromatographic or immunochemical methods such as ultraviolet-visible absorption spectrometry, absorption spectrometry in the near and middle infrared sections, fluorescence and chemiluminescence spectrometry, X-ray fluorescence spectrometry, atomic absorption spectrometry, atomic emission spectrometry (AES), inductively-coupled plasma atomic emission spectrometry (ICP-AES), nuclear magnetic resonance (NMR), mass spectrometry (MS) and mass spectrometry in tandem (MS/MS) (Mastovska, 2011).

Other separations methods are used in determined analysis such as capillary electrophoresis (CE), electro capillary chromatography (ECC) and polarimetry (Rouessac & Rouessac, 2003).

5. Veterinary drugs and growth promoters in food of animal origin

Currently, rearing animals aimed at feeding the human population mainly depends on using pharmacologically-active compounds. Using drugs in the animals is fundamental for animal health and wellbeing and for the economy of agribusiness. The reported benefits are mainly derived from keeping animals in good health, thereby reducing the possibility of a disease becoming transmitted from animals to humans. However, residues from drugs used in producing food of animal origin could increase the risk of disease in the people who consume products from treated animals.

In principle, all pharmaceutical preparations administered to animals producing foodstuffs can give rise to residues in edible tissue, milk or eggs. In addition to drug dose, residue

levels depend on withdrawal time. In spite of most drugs representing a relatively low risk for the general public, when used responsibly and in line with instructions approved by the laboratories making veterinary drugs, adverse reactions have been frequently reported for some compounds; these would include antibacterial, antihelminthic, anticoccidial and antiprotozoal drugs and growth promoters.

5.1 Antibacterial drugs

Residues from antibacterial drugs in food products of animal origin can represent a danger for consumers. The poisonous effects are not very probable since the residues are present in very low concentrations. Some substances must receive particular attention due allergic reactions. The main hazardous effect is likely to be the development of resistant bacterial strains following sub-therapeutic doses of antimicrobials being ingested; such resistance could be transferred to other bacteria. This could include resistance being transferred from non-pathogenic organisms to pathogenic ones which would then no longer respond to standard drug treatment (the Institute of Food Technologists [IFT], 2006). However, the generally used antibacterial drugs are presented in Table 1.

Differences in substitutions in the basic ring structures between the various aminoglycosides account for the relatively minor differences in antimicrobial spectra and resistance and toxicity patterns. Aminoglycosides given in therapeutic dosages mainly cause ototoxicosis, but may also cause nephrotoxicosis, allergy and neuromuscular disturbances.

Chloramphenicol (an antibacterial belonging to the amphenicol group) has been used in treatment and prophylactically in food-producing animals for several years now (i.e. poultry, calves, pigs, sheep and fish). Chloramphenicol's most serious toxic effect is bone marrow depression which is generally dose-related and reversible but can sometimes be fatal in patients who are probably genetically predisposed. A toxic syndrome has been reported in newborn infants receiving large doses of chloramphenicol which is characterised by vomiting, hypothermia, cyanosis and circulatory collapse followed by death; this syndrome rarely occurs in adults. Chloramphenicol may also cause neuritis, encephalopathy with dementia and ototoxicity; its use is restricted in many countries, while it is totally banned for use in food-producing animals within the European Union and the USA. Chloramphenicol and its metabolites could be genotoxic (Lozano & Arias, 2008).

Differences in substitutions in the basic ring structures between the various aminoglycosides account for the relatively minor differences in antimicrobial spectra and resistance and toxicity patterns. Aminoglycosides given in therapeutic dosages mainly cause ototoxicosis, but may also cause nephrotoxicosis, allergy and neuromuscular disturbances.

Chloramphenicol (an antibacterial belonging to the amphenicol group) has been used in treatment and prophylactically in food-producing animals for several years now (i.e. poultry, calves, pigs, sheep and fish). Chloramphenicol's most serious toxic effect is bone marrow depression which is generally dose-related and reversible but can sometimes be fatal in patients who are probably genetically predisposed. A toxic syndrome has been reported in newborn infants receiving large doses of chloramphenicol which is characterised by vomiting, hypothermia, cyanosis and circulatory collapse followed by death; this syndrome rarely occurs in adults. Chloramphenicol may also cause neuritis, encephalopathy with dementia and ototoxicity; its use is restricted in many countries, while it is totally

banned for use in food-producing animals within the European Union and the USA. Chloramphenicol and its metabolites could be genotoxic (Lozano & Arias, 2008).

Antibacterials	Aminoglicosides	Streptomycin, kanamycin, amikacin, neomycin, apramycin
	Amphenicols	Chloramphenicol, thiamphenicol, florfenicol
	Beta-lactams	Penicilins, cephalosporins
	Macrolides	Erythromycin, spiramycin, kitasamycin, josamycin, desmycosin, mirosamycin, tilmicosin, leucomycin, tylosin
	Nitrofurans	Furazolidone, nitrofurazone, furaltadone, nitrofurantoin
	Quinolones	Ciprofloxacin, danofloxacin, difloxacin, enrofloxacin, flumequine, marbofloxacin, norfloxacin, ofloxacin
	Sulphonamides	Sulfadiazine, sulfadimethoxine, sulfamethazine, sulfadoxine, sulfaethoxypyridazine, sulfaguanidine, sulfamerazine, sulfamethoxazole, sulfapyridine, sulfamethoxydiazine, sulfamethoxypyridazine, sulfamonomethoxine, sulfathiazole, sulfaquinoxaline
	Tetracyclines	Chlortetracycline, oxytetracycline, demeclocycline, doxycycline, methacycline, minocycline
Antihelmintics	Benzamidazoles	Thiabendazol, flubendazol, fenbendazol, mebendazol, albendazol, oxfendazol, febantel
	Imidazotiazoles	Levamisole
	Organophosphates	Haxolon, coumaphos, dichlorvos
	Tetrahydropyrimidines	Morantel, pyrantel
	Salicylanilides	Closantel, niclosamide, oxyclozanide, rafoxanide
	Sustituted phenols	Dichlorophen, hexachlorophen
	Macrocycliclactones	Abamectin, ivermectin, moxidectin
	Piperazinederivates	Piperazine, diethylcarbamazine
Antiprotozoals	Benzamides	Aklomide, nitromide, dinitolmide
	Carbanilides	Nicarbazin, imidocarb
	Nitroimidazoles	Ronidazole, dimetridazole, metronidazole, ipronidazole
	Polyether ionophore	Monensin, narasin, lasalocid, salinomycin, maduramicin
	Quinolonederivates	Buquinolate, decoquinate, methylbenzoquate
	Triazines	Clazuril, diclazuril, toltrazuril
Growth promoters	Antibiotics	Monensin, salinomycin, bambermycin, avilamycin
	Anabolic hormones	Estradiol-17, progesterone, testosterone
	Synthetic steroidal	Boldenone, chlormadinone acetate, ethylenestrol, fluoxymesterone, medroxyprogesterone acetate, megestrol acetate, methandienone, methylboldenone, methyltestosterone, drostanolone, norethandrolone, norgestomet, norgestrel, nortestosterone oxymetholone
	Organic arsenicals	Arsanilic acid
	Peptide antibiotics	Avoparcin, bacitracin, efrotomycin, enramycin, thiopeptin
	Quinoxaline-1,4-dioxides	Carbadox, olaquindox
	Beta-adrenergic agonists	Bambuterol, bromobuterol, carbuterol, cimaterol, clenbuterol, dobutamine, fenoterol, isoproterenol, mabuterol, mapenterol, metaproterenol, pirbuterol, ractopamine, reproterol, rimiterol, ritodrine, salbutamol

Table 1. Drugs administrated in treatment and prophylactically in food-producing animals

Penicillins have low toxicity; hypersensitivity reactions, especially skin rashes, are by far their most common adverse effects. Gastrointestinal disturbances including diarrhoea, nausea and vomiting may also sometimes appear. No teratogenic effects have been reported. Some studies have indicated that sensitive people have experienced allergic reactions, such as genera pruritis (itching), difficulty in swallowing and talking, dyspnoea,

dermatitis caused by contact and urticaria (hives) caused by consuming residues present in meat and/or milk (Medina et al., 2008). The allergic reactions caused by penicillin and its derivatives have been considered by the JECFA committee as being determinant factors for evaluating and establishing safe residue levels in foodstuffs. The adverse effects associated with cephalosporins are similar to those described for penicillins.

Lincomycine-macrolide is used for the initial treatment of mild to moderate staphylococcal infections in calves, sheep, goats and pigs and it is also added in feed for growth-promoting purposes. Lincomycine is reported to cause gastrointestinal disturbances including diarrhoea, vomiting and nausea which that may prove fatal. Other adverse effects include skin rashes, urticaria, polyarthritis, hepatic damage and haematological disturbances (WHO, 1989).

All nitrofurans have been widely used in the prophylactic and therapeutic treatment of infections caused by bacteria and protozoa in pigs, cattle, poultry, rabbits and fish. The use of nitrofurans in food-producing animals has been controversial because residues from these drugs may be mutagenic and tumorigenic. Toxicological studies have shown that nitrofurazone is a carcinogenic but not genotoxic agent, whereas furazolidone has exhibited both carcinogenic and genotoxic properties (WHO, 1993). The metabolites from nitrofurans can remain stored for weeks or m animal proteins, including eggs from farmyard birds, species in which this compound has been used as an anticoccidial. The systemic use of nitrofurans in food producing animals has thus been prohibited in the USA and Europe, (EMEA, 1997).

Quinolones are synthetic antibiotics which are very effective in combating various diseases in animal husbandry and aquaculture. The most frequent adverse affects of quinolone antibiotics most frequently occurring adverse affects are gastrointestinal disturbances including nausea, vomiting, diarrhoea, headache, visual disturbances and insomnia. Rashes, pruritus and epidermal necrolysis have sometimes also occurred (Jimenez et al., 2011).

The residuality of sulphonamides used in treating coccidial and bacterial infections and also as growth-promoting agents may cause hypersensitivity reactions, mainly cutaneous rash; however, no anaphylactic manifestations caused by this type of residues is known.

There is sufficient evidence indicating that ingesting the antibiotics in sub-therapeutic doses makes a significant contribution towards the appearance of resistant microorganisms in animals which can become transmitted to humans, thereby provoking difficult to treat infections. Some sample studies have found antibiotic-resistant coliform microorganisms in raw and cooked meat. Likewise, antibiotics consumed by human beings from residues present in food of animal origin lead to an alteration of intestinal flora and consequently a reduction of bacteria competing with pathogenous microorganisms, thereby increasing the risk of disease.

Tetracyclines can generate bacterial resistance; oxytetracycline particularly induces antibiotic resistance in coliform microorganisms present in the human intestine. Recognition of this effect has been used by the JECFA committee as the point of reference for defining acceptable consumption levels for different antibiotics.

The problem of resistance is not the only motive for the medical community's preoccupation. Farmers and veterinarians are worried because bacterial resistance in farm

animals is interfering with drug efficacy thereby leading to the use of greater concentrations than those initially established as being therapeutic. However, in spite of antibiotics being the type of veterinary drugs most used in the agribusiness industry, there are few options to choose from due to the limited offer of drugs which have been approved for use in animals producing foodstuffs compared to those regarding therapeutic use in humans.

5.2 Antihelminthic, anticoccidial and antiprotozoal drugs

Parasitic diseases constitute an ever present threat in rearing birds and livestock, but they may be controlled by adding low levels of drugs to daily rations. The drugs generally against internal parasites affecting animals collectively called helminths are shown in Table 1. Such drugs are used at levels which do not allow resistant strains to develop and also become rapidly metabolised in an animal's organism so that the residues in edible tissues are minimal.

Benzimidazoles, like thiabendazole, are used in sheep, cattle, horses, pigs and poultry. They become rapidly eliminated from the organism due to their high solubility; however, some studies has shown that these compounds are teratogenic and nephrotoxic in mice and ewes (Danaher et al., 2007). Mebendazole metabolits (hydroxylmebendazole and amino-mebendazole) belonging to the benzimidazole group and which are widely used as an antinematode in horses, sheep, pigs and poultry, have also been shown to have teratogenic effects (Buchmann et al., 1992).

Levamisol is the most well-known drug from the imidazothiazole group, which has a broad spectrum of activity against nematodes; however, it has been found that it induces idiosyncratic organulocytosis in some individuals. Levamisol's toxic effects have caused preoccupation in the regulatory bodies and, given that these effects the original compound than in its metabolites, this is the analyte of interest in tissue samples.

Organophosphates represent one of the alternatives for treating benzimidazole-resistant nematodes; haloxon (being one of them) is the safest and has been approved by the US Food and Drug Administration (FDA) for use in sheep, cattle and goats. By contrast, dichlorvos has an acceptable antihelminthic spectrum in cattle and sheep, but it does not have FDA approval for use in ruminants due to its suspected carcinogenic effects and narrow safety margin (Botsoglou & Fletouris, 2000).

Ivermectin, a macrocyclic lactone, is exceptionally effective in very low dosages against nematodes and arthropod parasites in cattle and has been widely used for treating endo- and ecto-parasites in cattle, sheep, goats and pigs; however, ivermectin has had a teratogenic effect in rats, rabbits and mice (Moreno et al., 2008).

Anticoccidial and antiprotozoal drugs are generally used in the poultry industry against protozoan infections caused by pathogenic species of *Eimeria.* Some compounds used as antibacterial drugs are also used as coccidiostats, including sulphaquinoxaline, sulphadimethoxine, sulphamethoxypyridazine, sulphachlorpyrazine, sulphamethazine, sulphaguanidine, furazolidone, nitrofurazone, tetracycline and chlortetracycline. Table 1 shows compounds whose primary function and use are related to antiprotozoal drugs.

A number of nitroimidazoles have already been banned within the European Union, even for therapeutic purposes, since they are mutagens and suspected carcinogens. The use of ronidazole has been banned by Council Regulation 3426/93/EEC (Official Journal of the European Communities, 1993) whereas dimetridazole use is banned by Council Regulation 1798/95/EEC (Official Journal of the European Communities, 1995). Their antibacterial and mutagenic activity is closely related to the reduction of the 5-nitro group, which is common to all nitroimidazole drugs. Metronidazole is used for treating bovine trichomoniasis by topical application or intravenous injection but it is a genotoxic carcinogen in animals.

Polyether antibiotics are produced by various actinomyces, mostly *Streptomyces* species, and constitute the agents most widely used by the poultry industry over the last two decades. They provide excellent disease control and are refractory for the development of resistance. They have a low therapeutic index but may be very toxic in certain species; salinomycin and narasin can be fatally toxic in turkeys, for example (Weissinger, 1994).

The problem of residues from antihelminthic, anticoccidial and antiprotozoal drugs may be easily controlled by imposing obligatory withdrawal times, generally 7-10 days, but this is unfortunately not always respected. On the other hand, given the large number of drugs which may be easily obtained on the market, many producers change one compound for another to avoid resistance becoming developed to drugs; however, this increases the degree of exposure to thema the misms, which may lead to yet another problem if it is taken into account that these drugs are also used as growth promoters.

5.3 Growth promoters

Growth promoters are substances which produce improvements in growth rate when added to animal feed in sub-therapeutic dosages over an extended period of time. Table 1 shows the compounds most commonly used for this purpose. The anabolic hormonal-type growth promoters can be classified according to their chemical structure or origin into endogenous sex steroids, steroidal compounds, not naturally occurring non-steroidal compounds and polypeptide hormones.

Anabolic hormones (estradiol-17 and progesterone - two female sex hormones, and testosterone - one male sex hormone) are used for increasing body mass in livestock rearing. Synthetic steroidal compounds have only been approved for therapy regarding reproductive behaviour and disorders in non-food-producing animals; however, they are used illegally around the world. Boldenone, chlormadinone acetate, ethylenestrol, fluoxymesterone, medroxyprogesterone acetate, megestrol acetate, methandienone, methylboldenone, methylthisterone, drostanolone, norethandrolone, norgestomet, norgestrel, norethisterone (nandrolone), norethisterone decanoate, oxymetholone, and stanozolol would be examples of synthetic steroidal compounds which have only been approved for therapy regarding reproductive behaviour and disorders in non-food-producing animals.

Zeranol and stilbene estrogens, including diethylstilbestrol, hexestrol and dienestrol, are the major non-steroidal not naturally occurring compounds included in the class of anabolic drugs and somatropin is the most common polypeptide compound affecting growth. Diethylstilbestrol, hexestrol and dienestrol are all stilbene estrogens which are currently banned worldwide for use in food-producing animals. They are genotoxic, not easily

metabolised compounds, which are considered capable of irreversibly initiating the carcinogenic process even in small residue concentrations. Diethylstilbestrol and hexestrol have been legally permitted for use as anabolics for quite some time in many countries, whilst the use of dienestrol, which is a diethylstilbestrol metabolite, has been restricted to illegal practice (Dickson, 2003).

Using these compounds, either natural or synthetic, as growth promoters in meat-producing animals has not been allowed in the European Union since 1988, due to potential adverse effects to human health, unlike in the United States where some anabolic hormonal-type growth promoters are permitted.

In vivo studies have demonstrated DNA strand breaks and oxidative damage being triggered by desencadenados por the 17-β estradiol, thereby leading to this hormone being considered as triggering a genotoxic effect (for example, the proliferation of carcinogenic mammary cells); however, the dosage at which these alterations occur is greater than that at which endocrine effects are produced in animals (Mikus et al., 2001).

Testosterone's adverse effects are due to its hormonal activity, particularly in the prostate gland. Testosterone is also considered to be potentially embryotoxic and its consumption in therapeutic doses has led to the induction of hepatic cystitis (Durlinger et al., 2002).

Following the ban of stilbene and other hormonal-type growth promoters, interest has focused on alternative compounds for promoting live weight gain in food-producing animals. The beta-adrenergic agonists constitute such group of compounds, clembuterol being the main one. It has been reported that consuming calf liver in Spain and France containing clenbuterol residues has induced muscular tremors, tachycardia, muscular pain, nervousness, headache, vertigo, nausea, vomiting and fever. It has also been used as an anabolizant steroid; clenbuterol is used as a tocolytic in cows, thereby supposing an additional risk.

A controversy has arisen around these events regarding whether to accept or prohibit using clenbuterol in animal production. This drug increases channel performance, it is not potentially oncogenic or mutagenic and is only embryotoxic in large doses whilst its adverse effects on consumers becomes evident when recommended withdrawal times are not respected and when excessive doses are used, whether through inadequate management or aimed at increasing animals' weight gain even more (Brambilla et al., 2007).

The foregoing has led to clenbuterol being a highly controlled drug today in many countries which have developed programmes and mechanisms for monitoring it and its follow-up. However, in spite of these controls and warning signs, unfortunate events involving adverse reactions continue to be presented, as happened in November 2005 in Jalisco, México, when about 225 people experienced trembling, headaches and discomfort after having consumed beef containing residues of this type (Gojmerac, 2002).

Arsanylic acid, peptide antibiotics and quinoxaline-1,4-dioxides are non-steroidal compounds used as growth promoters in different animal species. Arsanylic acid and its sodium salt are most commonly used, particularly in pigs. They are also efficacious in the egg-producing industry and were previously approved for use in egg-laying hens. However, their use in animals is generally rather limited and the risk–benefit ratio is questionable because these drugs can produce toxicosis, known as peripheral nerve demyelination.

In spite of the aforementioned effects, the Codex Alimentarius considers it unnecessary to establish an LMR for anabolic hormones as it is improbable that residues arising from the correct used of these substances as growth stimulators represent a danger for human health. It has also been demonstrated that the endogenous concentration of these hormones is greater when they are administered exogenously. Another reason negating the potential risk of this type of substance is the availability of metabolic routes which become rapidly degraded, meaning that the residues which the meat of treated animals may contain do not affect a consumer's endocrine system. However, dispositions in Europe regarding these substances are stricter and do not allow any residual level of anabolizant drugs in meats.

Peptide antibiotics are compounds usually containing D-amino acids. They are usually added to animal feeds in low concentrations and produce residues in tissues at very low or undetectable levels. Unfortunately, most peptide antibiotics' metabolic pathways have not yet been elucidated. These antibiotics are regulated under separate legislation within the European Union (Brogden et al., 2003).

Quinoxaline-1,4-dioxides and their possible residues in edible animal products have caused much debate regarding their mutagenic and carcinogenic potency. Carbadox was initially the main drug in use, but suspicion as to its safety arose because this compound exhibited both genotoxic and mutagenic activity. Olaquindox is also a strongly mutagenic agent but seemingly devoid of carcinogenic activity.

6. Toxins in food of animal origin

Toxins have a biological origin, mycotoxins, phycotoxins and phytotoxins having attracted most attention due to their potential residuality in foodstuffs, including animal subproducts.

6.1 Mycotoxins

Mycotoxins are secondary metabolites from fungi, mainly from the species *Aspergillus*, *Fusarium* and *Penicillium*, aflatoxins, ochratoxins, zearalenone, trichothecenes and fumonisins having been the most studied to date. The foodstuffs fundamentally contaminated by these toxins are grains and cereals constituting the main source of contamination for human beings. However, farm animals consuming contaminated foodstuffs may generate residues in meat, viscera, milk and eggs. Residuality is determined by contamination by high concentrations in foodstuffs ingested by animals, this being very uncommon, and also by theway in which the xenobiotic becomes metabolised in the organism. Mycotoxins do not become totally destroyed during cooking or industrialisation of foodstuffs due to their heat-stability.

The types of mycotoxicosis (disease resulting from consuming mycotoxins) in human beings are mainly chronic. These would include Balkan endemic neuropathy in Russia caused by the consumption of ochratoxin A which generates nephrotoxicosis, alimentary toxic aleukia in the former Soviet Union associated with dermatitis, vomiting and hematopoietic alterations caused by trichothecenes (diacetoxiscirpenol and T-2 toxin), possible endocrinal alterations related to reduced masculine fertility caused by consuming zearalenone (such toxin acting as an xenoestrogen), hepatic cancer caused by aflatoxin B1 and possible esophagic and renal cancer caused by fumonisin and ochratoxin A, respectively. The IARC

has classified aflatoxin B1 within group 1 (proven carcinogenic effect on humans) and fumonisin B1 and ochratoxin A within group 2B (possibly carcinogenic to humans) (IARC, 1993). The evidence from *in vitro* studies has shown that zearalenone is a probably implicated in cancer of the reproductive system (Khosrokhavar et al., 2009).

Aflatoxins and ochratoxin A are the main mycotoxins which can generate residuality and attention concerning them as being animal subproduct contaminants has mainly been focused on their presence in milk; however, it has been demonstrated that they can also generate residuality in meat and eggs.

6.1.1 Aflatoxins

Aflatoxins (AF) B1 B2 G1 and G2 are produced by fungi from the genera *Aspergillus*. AFB1 may be bioactivated through CYP450 enzymes to become an epoxide which is able to form adducts with DNA, meaning that it has been considered that AFB1 undergoes bioactivation in the organism. AFB1 may also become hydroxylised to AFM1 and be excreted in milk. It has been estimated that 1% to 6% of AFB1 ingested by a milk-producing cow could be excreted as AFM1 in milk, depending on bovine productivity. AFM1, like AFB1, may form an epoxide and alter DNA sequences. IARC is considered to be an AFM1 in group 2B (IARC, 1993). MRL regulated in different countries ranges from 0.05 to 0.5 μg/L; MRL has also been established for AF consumed by ruminants (FAO, 2003). Experimental studies have shown that when animals consume foodstuffs contaminated by high levels of AF, that it is difficult to find naturally, the liver and kidneys are the organs where most toxins become accumulated, and their presence in muscle is scarce (Bailly & Guerre, 2009). These types of studies have also demonstrated the presence of AF in eggs from different avian species.

6.1.2 Ochratoxin A

Ochratoxin A (OA) is produced by fungi from the genera *Aspergillus* and *Pencillium*, the former being from tropical regions and the latter from temperate regions. OA may thus be widely distributed throughout the world. OA may become biotransformed through hydrolysis reactions in which metabolites become less toxic by the opening of the lactone ring which occurs during bioactivation. Detoxification may occur in ruminants through digestive flora action before absorption, thereby limiting the possibility that OA might be found in milk and/or beef (Bailly & Guerre, 2009). However, a recent study evaluating the presence of OA in cows' milk formulas for infants found contamination in 72% of the samples analysed, levels around 690 ng/L being found (Meucci et al., 2010). It has been shown that OA may become accumulated in pigs' kidneys. In countries such as Denmark, OA levels in these organs are regulated since porcine ochratoxicosis is common.

6.1.3 Fusariotoxins

The fusariotoxins are mycotoxins which are produced by fungi from the genera *Fusarium*, zearalenone (ZEA), the fumonisins (FUM) and the trichothecenes (TCT) being the most important for public health.

ZEA is frequently implicated in reproductive disorders in animals and occasionally in hyperestrogenism syndromes in humans. ZEA becomes biotransformed in the intestine by the mucosa or bacterial flora and involves the formation of α- and β-zearalenol and α- and β-zearalanol. Alpha-zearalanol and β-zearalenol have greater estrogenic power than ZEA since they bind with greater force to their corresponding receptors (Zinedine et al., 2007). Alpha-zearalanol has been employed as growth promoter in cattle. Studies orientated towards determining residuality through experimentation have suggested that residues are not present in meat or eggs, even at high doses. However, a recent study has show that the presence of α-zearalenol in meat-based foodstuffs for infants reached levels of 30.5 μg/kg; the same study demonstrated the presence of mycoestrogens (ZEA, α-zearalenole and β-zearalenole) in infants' cow milk formulas (Meucci et al., 2011).

FUM properties suggest that their presence in animal meat does not represent an important source of contamination, since they are poorly absorbed. FUM produce liquefaction of the brain in horse and pulmonary oedema in pigs; ruminants and birds are more resistant. They have been correlated with oesophageal cancer in humans in some parts of the world and it has also been presumed that they may cause neural tube alterations. Their presence has been demonstrated and in the liver and kidneys of turkeys fed with the maximum levels permitted in Europe (Tardieu et al., 2008).

The main TCT of interest in producing animals are T-2 toxin, HT-2 toxin, diacetoxiscirpenol and deoxinivalenol. The TCT do not usually represent a risk of contamination in food of animal origin due to their rapid metabolism (Bailly & Guerre, 2009).

6.2 Phycotoxins

Around 75 species of marine micro-algae, belonging to the dinoflagellate group, produce secondary metabolites which represent potent toxins, generically called phycotoxins. These organisms form part of the marine plankton and therefore the aquatic food chain leading to filtrator mollusks, gastropods, crustaceans and fish which can accumulate toxins being consumed (FAO, 2004).

The microalgae population may increase suddenly and generate an algal bloom which has increased in frequency, intensity and geographical distribution during recent years. Amongst the explanations put forward to explain for this phenomena has been the increased use of coastal waters for aquaculture, eutrophication caused by domestic, industrial and agricultural residues, the mobility of trace metals and humic substances due to deforestation and/or acid rain and changes in climatic conditions (Erdner et al., 2008). Reports of phycotoxins poisoning have increased during the last few years, perhaps due to the scientific community's greater knowledge and interest in the matter or due to the increase in algal bloom. Such poisoning is mainly acute course; however, there is interest in evaluating the effects being triggered by chronic consumption. The lack of studies on animals which are continuously exposed to phycotoxins and the scarce availability of certified reference materials have led to difficulties in evaluating risk, developing analytical methodologies and regulating these substances.

Studying toxins produced by algae has classically been approached according to the type of poisoning which they have caused. Four groups of toxins can thus be distinguished, causing paralytic shellfish poisoning, diarrhoeic shellfish poisoning, amnesic shellfish

poisoning and neurotoxic shellfish poisoning (FAO, 2004). Another group of phycotoxins of interest due to their accumulation in sea fish are the ciguatoxins causing ciguatera fish poisoning.

6.2.1 Paralytic shellfish poisoning

The toxins responsible for this poisoning are mainly produced by algae from the genera *Alexandrium, Gymnodinium* and *Pyrodinium*. Chemically, they correspond to tetrahydropurin molecules, saxitoxin (STX) having the most importance. PSP incidence and geographical distribution has increased since the 1970s; this poisoning was initially confined to temperate waters in Europe, North America and Japan, but is now considered to be worldwide problem (FAO, 2004).

STX becomes rapidly absorbed through the gastrointestinal tract and equally has rapid distribution, metabolism and excretion. STX selectively blocks (and with high affinity) sodium-dependent channels present in nerves, skeletal muscular fibres and most cardiac muscular fibres, thereby reducing or eliminating the action of propagation potential (Etheridge, 2010).

Paralytic shellfish poisoning symptoms begin in human beings within the first 30 minutes following consumption of contaminated foodstuff and the onset of numbness and/or pins and needles around the lips, gradually extending to the face, neck, arms and legs. Headaches, nausea, lack of muscular coordination and, occasionally, temporary blindness occur. There may be paraesthesia in the arms and legs, motor inability and difficulty in talking in moderate cases and paralysis of respiratory muscles leading to death may occur in severe cases.

6.2.2 Diarrhoeic shellfish poisoning

Diarrhoeic shellfish poisoning toxins are produced by dinoflagellates from the genera *Dinophysis* and *Protoceratium*, having worldwide distribution. Okadaic acid (OA) and its analogues (dinophysis toxins , pectenotoxin and yessotoxin are included within this group. However, yessotoxin and pectenotoxin produce different toxicological effects in experimental animals. Yessotoxin is related to lesions in the cardiac muscle, liver, pancreas and cerebral neurons and pectenotoxin is clearly hepatotoxic. The effect which yessotoxin and pectenotoxin may have on human beings remains unknown (Dominguez et al., 2010).

OA was initially reported in Japan and Europe, areas in which diarrhoeic shellfish poisoning has had greater importance. OA and its analogues are potent protein phosphatase inhibitors (serine/threonine phosphatases PP1 and PP2A) which dephosphorylate molecules closely related to metabolic processes. It has been postulated that OA induces diarrhoea due to an alteration in hydric balance in the intestines via one of the two following mechanisms: stimulating the phosphorylation of proteins controlling sodium secretion in enterocytes or promoting the phosphorylation of intercell binding proteins regulating solute permeability. Diarrhoeic shellfish poisoning is characterised by diarrhoea, nausea, vomiting and, in some cases, abdominal pain which can begin within the first 3 or 12 hours after having consumed contaminated organisms. No lethal effects have been reported concerning human as having been caused by OA or its analogues.

6.2.3 Amnesic shellfish poisoning

This poisoning is also known as domoic acid (DA) poisoning since memory loss is not always present. It was described for the first time in Canada (Prince Edward Island) in 1987 when 105 people became poisoned after consuming blue mussels. There have also been several reports of poisoning involving effects on wild life, demonstrating that the toxin forms part of the food chain; the toxin responsible for this has been DA which is produced by diatoms from the genera *Pseudo-nitzschia*.

The DA mechanism of action acts on excitatory amino acid receptors (L-glutamate, L-aspartate) and/or synaptic transmission. DA activates specific excitatory amino acid L - glutamate receptors producing an excessive accumulation of calcium resulting in cell death. The kainate receptor is DA's primary target. Recent interest in DA has been centred on recognising that effects can result following chronic exposure to it at low concentrations, given the discovery that chemical route alteration leads to neurological disturbances.

Intestinal absorption is limited (5%-10% of the dose administered to experimental animals). It has high distribution in the blood compartment and scarcely penetrates the hematoencephalic barrier. There is no evidence that DA may become metabolised. Elimination occurs via the kidneys. Poisoning in humans produces gastroenteritis which may be accompanied by headache, confusion and permanent loss of short-term memory (FAO, 2004).

6.2.4 Neurotoxic shellfish poisoning

Neurotoxic shellfish poisoning, which is endemic on the Gulf of México and the eastern coast of Florida, is caused by brevetoxin (BTX) produced by the dinoflagellate *Gymnodinium breve* (synonyms: *Ptychodiscus breve, Karenia brevis*) present in red-tides. This alga has the special feature of being able to form aerosols due to wave action thereby constituting a risk of aerial exposure.

BTX are liposolube toxins consisting of around 14 different substances, leading to depolarisation opening sodium channels in cell membranes and increasing the inflow of sodium causing persistent and repetitive activation. Symptoms caused by oral exposure to BTX occur within the first 30 minutes to 3 hours after consuming contaminated organisms and include vomiting, diarrhoea, shivering, sweating, conflicting perception of temperature, hypotension, arrhythmias, numbness, paraesthesia of the lips, face and extremities, cramps, bronchoconstriction, paralysis, convulsions and coma. There have been no reports of lethality. Respiratory difficulty and irritation of the mucosa are the most common symptoms when inhalatory exposure occurs.

6.2.5 Ciguatera fish poisoning

Ciguatera fish poisoning is caused by ciguatoxin (CTX) which is produced by dinoflagellates from the genera *Gambierdiscus.* CTX becomes accumulated through the food chain, from small herbivorous fish up to large carnivorous fish. This poisoning has passed from being a problem limited to insular regions which affected local communities to being a global matter, given the worldwide consumption of seafood and international tourism. This is the most common poisoning caused by seafood and may affect up to 50,000 people annually

(FAO, 2004). CTX are liposoluble toxins having 13 to 14 rings fused in a rigid structure. CTX bind to sodium channels causing them to open during cell membranes' potential repose altering bioenergetic mechanisms. CTX acts on the same receptor as BTX does but with greater affinity (Lehane & Lewis, 2000).

Gambierdiscus toxicus, the alga specie most commonly related to CTX production, is distributed throughout the tropical region of the Pacific Ocean, western Indian Ocean and the Caribbean where CFP is endemic. Many coral fish species are involved, including herbivores and carnivores. The latter constitute the main vector for poisoning in humans, particularly *Muraenidae* (moray eels), *Lutjanidae* (snappers), *Carrangidae* (carrangs), *Scombridae* (mackerels) and *Sphyraenidae* (barracuda) (FAO, 2004).

CTX become rapidly absorbed through the intestine and are mainly excreted in the faeces via the bile. The symptoms are gastrointestinal or neurological in nature, the former include vomiting, diarrhoea, nausea and abdominal pain. The neurological symptoms which can begin later include pins and needles in the lips, hand and feet, disturbances in perception of temperature, severe pruritus and fatigue. Some patients can experience pain (muscular, articular and dental) and anxiety. There may be hypotension and bradycardia in severe cases and death may occur, though this is not very common. The neurological symptoms can persist for years in some cases; it seems that the toxin may become accumulated in fat and be released in certain circumstances or also produce an immunological response (Shoemaker et al., 2010).

Other phycotoxins of interest due to their residuality are the azaspiracida, discovered in 1998, and whose symptoms resemble those of diarrhoeic shellfish poisoning; the cyclic imins (gymnodimine, spirolids, pinnatoxins, prorocentrolide, spirocentrimine and pteriatoxin), still have no associated effects in humans, but their residuality does interfere with the analytical methodologies used for determining the presence of marine toxins and cianotoxins (nodularin and cylindrospermin) by producing hepatoxicity and inhibiting protein synthesis.

6.3 Phytotoxins

Plants have secondary metabolites with which they defend themselves from the aggression of herbivorous animals. Many of these are toxic for humans and animals, causing numerous pathologies. Given that the diverse compounds present in plants are degraded during digestion and/or the metabolism of xenobiotics in animals of livestock interest, only some manage to contaminate products of animal origin. Milk is the main subproduct which has been studied in which toxins from plants may be present; however, their presence has also been demonstrated in muscle, viscera and eggs. Regulations regarding these toxins are scarce and are mainly orientated towards their presence in botanical products having pharmacological uses; however, there is growing interest in making advances in this field. The regulating entities have shown the greatest interest in pyrrolizidine alkaloids amongst the plant toxins due to their abundance and proven toxic effects.

6.3.1 Pyrrolizidine alkaloids

Pyrrolizidine alkaloids (AP) are present in a large variety of plants and are perhaps the most widely distributed toxins. Foodstuffs or botanically-based remedies represent the most

probable risk of exposure for humans; however, food of animal origin can also contain pyrrolizidine alkaloids. It has been presumed that more than 6,000 plant species contain AP, mainly those belonging to the families *Asteraceae* or *Compositae* (genera *Senecio* and *Eupatorium*), *Boraginaceae* (genera *Heliotropium* and *Echium*) and *Fabaceae* or *Leguminosae* (genera *Crotalaria*).

APs are heterocyclic compounds which are mainly derived from four necine bases: retronecine, heliotridine, otonecine and platynecine. AP can become enzimatically hydrolised or oxidised; the resulting N-oxide is slightly toxic and may also be found in plants. AP becomes bioactivated to a toxin pyrrole through CYP450 which is electrophilic and unstable and reacts rapidly with endogenous macromolecules (particularly with DNA forming adducts). DNA adducts could be a continuous source of carbon ions originating new adducts, meaning that the total elimination of AP derivates may take months (Edgar et al., 2011).

AP levels in foodstuffs are rarely so significant that they can cause acute diseases; however, low levels and continuous exposure could lead to the presentation of chronic diseases which could hardly be attributable to the toxin in foodstuffs. Pyrroles cause thickening and occlusion of the hepatic vessels resulting in veno-occlusive disease and cirrhosis. They can also alter the pulmonary vessels, causing pulmonary hypertension and congestive cardiac failure. The IARC has classified lasiocarpine, monocrotaline and riddelliine AP within group 2B (probably carcinogenic for humans, given the pertinent evidence regarding how they are cancerigens in animals) (Edgar et al., 2011).

It has been demonstrated that 0.1% to 4% of AP in foodstuffs for lactating cows and sheep could be excreted in milk (Hoogenboom, 2011). The meat and viscera of cattle fed on AP-rich plants may contain the toxin and its derivates at levels reaching 250 mg/kg in muscle and 2,500 mg/kg in liver (Fletcher et al., 2011). AP can potentially be present in eggs when fowl are fed on AP-rich diets, thereby constituting a risk for human health (Eröksüz et al., 2008). In spite of potential contamination in milk, meat, viscera and eggs, honey is the only foodstuff of animal origin which has been shown to be naturally contaminated with AP; a large number of plants habitually used in apiaries could be a source of significant levels of the toxins.

Other alkaloids of interest due to their potential residual effect on animal subproducts are the indolizidinics (especially swainsonine) producing lysosomal storage disease and piperidine (coniin and gamma-conicein) which in acute form can produce muscular paralysis and (chronically) teratogenesis (Panter & James, 1990). These last named alkaloids seem to be responsible for coturnism, a human disease which occurs following the consumption of migrating wild European quail; it is characterised by weakness, muscular pain, paralysis of the legs, vomiting and myoglobinuria. It has been postulated that these symptoms occur due to the accumulation of coniin in birds' tissues following the consumption of *Conium maculatum* (López & Bianchini, 1999).

6.3.2 Glucosinolates

Glucosinolates, known as mustard oil glucosides as they confer the characteristic flavour on black mustard (*Brassica nigra*), are secondary metabolites produced by plants belonging to the order *Brassicales*, mainly in the family *Brassicaceae (*or *Cruciferae)*. Many plants which are

common in the human diet belong to this family (broccoli, Brussels sprouts, cabbage and cauliflower) and several genera (*Brassica*, *Crambe*, *Sinapis* and *Raphanus*) including crops used for producing vegetal oils. The plant *Camelina sativa* (false flax) has recently aroused interest due to biofuel production. When the oils are extracted, the glucosinolates remain in the seed; the resulting press cake is used in animals diet (EFSA, 2008).

The glucosinolates become hydrolysed by enzymatic action giving place to isothiocyanates, thiocyanates, oxazolidinethiones and nitriles. The thiocyanates interfere with iodine capture and the oxazolidinethiones with thyroid (T_3 and T_4) hormone synthesis, leading to hypothyroidism and thyroid gland enlargement. Consequently, metabolism in all tissues, including the reproductive organs, may become affected.

It has been shown that high glucosinolate consumption in lactating cows reduces iodine levels and increases thiocyanates in milk, the liver and the kidneys. The residues in milk account for around 0.1% of the dose received by animals; the residues in muscles and viscera are even lower. Residuality has also been found in eggs after rapeseed has been administered to egg-laying birds; these also acquire a disagreeable flavour (EFSA, 2008).

6.3.3 Ptaquiloside

Bracken fern (genus *Pteridium)*, considered one of the five most abundant plants in the world, contains a norsesquiterpene glucoside called ptaquiloside. It has been proved that ptaquiloside may cause tumours in the urinary bladder, mammary glands, intestine and other organs in laboratory rodents. It causes degeneration of the retina in sheep and causes urinary bladder cancer known as bovine enzootic haematuria in cattle. There is epidemiological evidence relating the consumption of bracken fern in humans (Japanese population) suffering from esophagic and gastric cancer, possibly caused by ptaquiloside. Around 8.6% of the ptaquiloside present in *P. aquilinum*, consumed by lactating cows, is excreted in milk, thereby making contaminated raw milk a risk for human health (Alonso-Amelot et al., 1996).

6.3.4 Tremetol

This is a liposoluble compound mixture of terpene, sterols, tremetone, hydroxytremetone and dehydrotremetone; the last three are present in a ketonic fraction. It is present in the perennial plants white snakeroot (*Ageratina altissima*, previously called *Eupatorium rugosum*) and rayless goldenrod (*Haploppapus heterophyllus*). This poisoning reached an epidemic proportion during the 18th and 19th centuries in the USA (Indiana, Illinois and Ohio), producing high mortality, and its aetiological agent was only discovered in 1910. Cows accumulate the toxin in fat and excrete it in milk. The toxin becomes diluted in milk reception tanks, making such poisoning not very common in modern milk production systems. Tremetol produces acidosis, hyperglycaemia and ketonaemia, as a consequence in the Krebs cycle inhibitor. The symptoms of poisoning are anorexia, listlessness, weakness, stiffness of muscles, vomiting, constipation and coma. Marked acidosis and ketosis may lead to death (Lewis & Elvin-Lewis, 2003).

7. Conclusions

Interest in diseases caused by food has mainly been orientated towards the acute presentation principally produced by microbiological agents; however, consuming food

contaminated by chemical substances could lead to chronic exposure leading to the presentation of diseases lacking an apparent cause and being difficult to diagnose. Foods of animal origin presuppose the risk of contamination, whether from drugs and growth promoters used for optimising livestock production systems, or with biological toxins present in food ingested by animals. It is thus necessary to control these substances in foods, thereby supposing technological and institutional efforts.

Sanitary authorities must thus promulgate and ensure compliance with standards and guidelines concerning the production of harmless foodstuffs. Achieving such objective represents a great challenge for underdeveloped and developing countries due to institutional difficulties and the limited availability of equipment and qualified personnel. All nations must make it a priority to try to ensure the safe consumption of foodstuffs by their populations, exercising strict sanitary control aimed to avoid problems of health in the population and preventing the appearance of new problems affecting the development of the agro-food industry and global trade in foodstuffs.

8. References

Alonso-Amelot, ME.; Castillo, U.; Smith, BL. & Lauren, DR. (1996). Bracken ptaquiloside in milk. *Nature*, Vol. 382, No.6592, pp 587, ISSN 0028-0836

Asociación Española de Farmacéuticos de la Industria. (March 2001). *Validación de Métodos Analíticos* (First Edition), Monografías de A.E.F.I., ISBN 84-89602-33-6, Barcelona, España

Bailly, JD. & Guerre P. (2009). Mycotoxins in Meat and Processed Meat Products In: *Safety of meat and processed meat,* Fidel Toldrá, pp (1-699) Springer, ISBN 978-0-387-89025-8, New York, USA

Botsoglou, N. & Fletouris, D. (2000). *Drug Residues in Foods, Pharmacology, Food Safety and Analysis,* (First Edition), Marcel Dekker Inc., ISBN: 0-8247-8959-8, New York, USA

Blondin, P. & Sirard, M. (1998). Oocyte quality and embryo production in cattle. *Canadian* 27 *Journal of Animal Science,* Vol. 78, (October 1998), pp. 513–516, ISSN 0008-3984

Brambilla, G.; Di Beza, S.; Pietraforte, D.; Minetti, M.; Campanella, L. & Loizzo, A. (2007). *Ex vivo formation* of gastric metabolites of clenbuterol: Preliminary characterisation of their chemical structure. *Analytica Chimica Acta*, Vol. 586, (July 2006), pp. 426–431, ISSN 0003-2670

Brogden, KA.; Ackermann, M.; McCray, PB Jr. & Tack, BF. (2003). Antimicrobial peptides in animals and their role in host defences. *International Journal of Antimicrobial Agents*, Vol. 22, (February 2003), pp. 465-478, ISSN 0924-8579

Buchmann, K.; Roepstorff, A. & Waller, P.J. (1992). Experimental selection of mebendazole resistant gill monogeneans from European eel *Anguilla anguilla*. *Journal of Fish Diseases,* Vol. 15 No. 5, (September 1992), pp. 393-400, ISSN 1365-2761

Danaher, M.; De Ruyckb, H.; Crooks, S.; Dowling, G. & O'Keeffe, M. (2007). Review of methodology for the determination of benzimidazole residues in biological matrices. *Journal of Chromatography B*, Vol. 845 No. 1, (July 2006), pp. 1–37, ISSN 1570-0232

Dickson, L. C.; Macneil, J.D.; Reid J. & Fesser A. (2003). Validation of Screening Method for Residues of Diethylstilbestrol, Dienestrol, Hexestrol, and Zeranol, in Bovine Urine Using Immunoaffinity Chromatography and Gas Chromatography/Mass

Spectrometry. *Journal of AOAC International*, Vol. 86, No. 4, (April 2003), pp. 631-639 ISSN 1060-3271

Dominguez, HJ.; Paz, B.; Daranas, A; Norte, M.; Franco, J. & Fernández,JJ. (2010) Dinoflagellate polyether within the yessotoxin, pectenotoxin and okadaic acid toxin groups: Characterization, analysis and human health implications. *Toxicon,* Vol. 56, No.2, (August, 2010), pp 191–217, ISSN 1879-3150

Durlinger, A.; Visser, J. & Axel T. (2002). Regulation of ovarian function: the role of anti-Müllerian hormone. *Reproduction,* Vol. 124, (April 2002), pp. 601–609 ISSN: 1470-1626

Edgar, JA.; Colegate, SM.; Boppré, M. & Molyneux RJ. (2011). Pyrrolizidine alkaloids in food: a spectrum of potential health consequences. *Food Addittives and Contaminants. Part A chemistry, analysis, control, exposure and risk assessment*, Vol.28, No.3, (March 2011), pp 308–324, ISSN: 1944-0049

Erdner,DL.; Dyble, J.; Parsons,ML.; Stevens, RC.; Hubbard K.; Wrabel, ML.; Moore, S.; Lefebvre, K.; Anderson, D.; Bienfang, P.; Bidigare, R.; Parker, MS.; Moeller, P.; Brand, L. & Trainer VL.(2008). Centers for Oceans and Human Health: a unified approach to the challenge of harmful algal blooms. *Environmental Health : a global access science source*,Vol.7, Suppl.2, (November, 2008), ISSN 1476-069X

Etheridge, SM. (2010). Paralytic shellfish poisoning: Sea food safety and human health perspectives. *Toxicon,* Vol. 56, No.2, (August, 2010), pp 108–122, ISSN 1879-3150

Eroksuz, Y.; Ceribasi, A.; Cevik, A.; Eroksuz, H.; Tosun, F. & Tamer, U. (2008). Toxicity of *Heliotropium dolosum*, *Heliotropium circinatum*, and *Senecio vernalis* in Parental Quail and Their Progeny, with Residue Evaluation of Eggs. *Turkish Journal of Veterinary and Animal Sciences*, Vol.32, No. 6, pp 475-482, ISSN 1303-6181

European Food Safety Authority, EFSA. (2008). Opinion of the Scientific Panel on Contaminants in the Food Chain on a request from the European Commission on glucosinolates as undesirable substances in animal feed, *The EFSA Journal* 590, pp 1-76, ISSN 1831-4732

European Agency for the Evaluation of Medicinal Products, EMEA. (1997). In Furazolidone, Summary Report, Committee for Veterinary Medicinal Products, EMEA/MRL, London, UK

Fletcher, MT.; McKenzie, RA.; Reichmann, KG. & Blaney, BJ. (2011) Risks from plants containing pyrrolizidine alkaloids for livestock and meat quality in Northern Australia. In: *Poisoning by plants, mycotoxins and related toxins,* Riet-Correa, F.; Pfister, J.; Schild, AL. and Wierenga, TL. pp (1-660) CABI Publishing, ISBN 9781845938338, USA.

Food and Agriculture Organization of the United Nations, FAO. (2003). *Worldwide regulations for mycotoxins in food and feed in 2003*. FAO Food and nutrition paper 81. ISBN 92-5-105162-32004, Rome, Italy

Food and Agriculture Organization of the United Nations. (2004). *Marine biotoxins*, FAO Food and Nutrition Paper 80, ISNN 0254-4725 Rome, Italy

Gojmerac, T.; Mandi, B.; Pleadin, K. & Mitak, M. (2002). Determination of Clenbuterol in Pig Liver Following Prolonged Administration of a Growth-Promoting Dose. *Food Technology and Biotechnology,* Vol. 40, No. 4, (November 2002), pp. 343–346, ISSN 1330-9862

Hoogenboom, LA.; Mulder, PP.; Zeilmaker MJ.; Van Den Top HJ.; Remmelink, J.; Brandon, EF.; Klijnstra, M.; Meijer, GA.; Schothorst, R. and Van Egmond, HP. (2011). Carry-over of pyrrolizidine alkaloids from feed to milk in dairy cows. Food Additives and Contaminants. Part A Chemistry, analysis, control, exposure and risk assessment, Vol.28, No.3, (March, 2011), pp 359-372, ISSN 1944-0049

International Agency for Research on Cancer, IARC. (1993). *Some Naturally Occurring Substances: Food Items and Constituents, Heterocyclic Aromatic Amines and Mycotoxins*. IARC Monographs on the evaluation of carcinogenic risk to humans 56. IARC, pp. (1- 599) ISBN 92 832 1256 8 Lyon, France

Jiménez, V.; Companyó, R. & Guiteras, J. (2011). Validation of a method for the analysis of nine quinolones in eggs by pressurized liquid extraction and liquid chromatography with fluorescence detection. *Talanta*, Vol. 85 No. 1, (July 2011), pp. 596-606. ISSN 0039-9140

Khosrokhavar, R.; Rahimifard, N.; Shoeibi S.; Pirali, M. & Hosseini, M. (2009). Effects of zearalenone and *a*-Zearalenol in comparison with Raloxifene on T47D cells. *Toxicology mechanisms and methods,* Vol. 19, No. 3, (March, 2009), pp 245-250, ISSN 1537-6516

Lehane, L. & Lewis, RJ. (2000). Review Ciguatera: recent advances but the risk remains. *International Journal of Food Microbioly,* Vo.61, No.2-3, (November, 2000) pp 91-125, ISSN 1879-3460

Lehman-McKeeman, LD. (2008). Absorption, distribution and excretion of toxicants, In: *Casarett &Doul's Toxicology. The basic science of poisons 7h Edition*, Curtis Klaassen, pp. (1-1309) McGraw-Hill, ISBN 978-0-07-147051-3, New York, USA

Lewis, WH. & Elvin-Lewis, PF. (2003). *Medical botany: plants affecting human health. 2nd Ed.* Wiley Press, pp (1-832), ISBN: 978-0-471-62882-8

LoBrutto, R. & Patel, T. (2007). Method Validation, In: *HPLC for Pharmaceutical Scientists,* Y. Kazakevich & R. Lobrutto, pp. 455-502, Wiley Interscience, ISBN-13: 978-0-471-68162-5, New Jersey, USA

Lodovico, C.; Marinovicha M. & Lotti M. (2008). Is the acceptable daily intake as presently used an axiom or a dogma?. *ToxicologyLetters*, Vol.180, No.2, (August, 2008), pp 93-99, ISSN 0378-4274

López, TA. & Bianchini, ML. (1999). Biochemistry of hemlock (*Conium maculatum* L.) alkaloids and their acute and chronic toxicity in livestock. A review. *Toxicon,* Vol.37, No.6, (June, 1999), pp 841-865, ISSN 1879-3150

Lozano, MC., & Arias, DC. (2008). Residuos de fármacos de origen animal: panorama actual en Colombia. *Revista Colombiana de Ciencias Pecuarias*, Vol. 21 No. 1, (March 2008), pp. 121-135, ISSN 0120-0690

Mastovska, K. (2011). Multiresidue analysis of antibiotics in food of animal origin using liquid chromatography-mass spectrometry. *Methods in Molecular Biology*, Vol. 747, (May 2011), pp. 267-307, ISSN 1940-6029

Medina, M.; González D. & Ramírez A. (2008). Detection of antimicrobial residues in animal tissues and tetracyclines in bones of pigs. *Revista de Salud Animal*, Vol. 30 No. 2, (May 2008), pp. 110-115, ISSN 0253-570

Meucci, V.; Razzuoli, E.; Soldani, G. & Massart, F. (2010). Mycotoxin detection in infant formula milks in Italy. *Food Addittives and Contaminants. Part A chemistry, analysis,*

control, exposure and risk assessment, Vol. 27, No. 1, (January, 2010), pp 94-71, ISSN: 1944-0049

Mikus, JH., Duff, GC., Krehbie, C.; Hallford, DM.; Walker, DA.; Graham, JD., & Ralphs, M H. (2001). Effects of an Estradiol Implant on Locoweed Consumption, Toxicity, and Recovery in Growing Beef Steers, *Professional Animal Scientist,* Vol. 17, No. 2, (June 2001), pp. 109-11, ISSN 1080-7446

Moreno, L.; Alvarez, L.; Ceballos, L.; Sánchez Bruni S. & Lanusse C. (2008). Pattern of ivermectin (sheep) and doramectin (cattle) residues in muscular tissue from various anatomical locations. *Food Additives & Contaminants: Part A*, Vol. 25, No. 4, (April 2008), pp. 406-412, ISSN 1944-0049

Panter, KE. & James LF.(1990). Natural plant toxicants in milk: a review. *Journal of Animal Science*, Vo.68, No.3, (March, 1990), pp 892-904, ISSN 0021-8812

Rouessac, F. & Rouessac, A. (2003). *Análisis Químico. Métodos y Técnicas Instrumentales Modernas* (First edition), McGraw Hill, ISBN: 9788448137854, Madrid, España

The Institute of Food Technologists. (2006). Comprehensive reviews Food Science and Food Safety, *Institute of Food Technologist*,Vol. 5. No. 3 (August 2006) pp. 71-137 ISNN 1541-4337

Shoemaker, RC.; House, D. & Ryan, JC. (2010). Defining the neurotoxin derived illness chronic ciguatera using markers of chronic systemic inflammatory disturbances: a case/control study. *Neurotoxicology and Teratology*, Vol.32, No.6, (December, 2010), pp 633-639, ISSN 1872-9738

Tardieu, D.; Bailly, JD.; Skiba, F.; Grosjean, F. & Guerre, P. (2008). Toxicokinetics of fumonisin B1 in turkey poults and tissue persistence after exposure to a diet containing the maximum European tolerance for fumonisins in avian feeds. *Food and Chemical Toxicology*, Vol. 4, No. 9, (January, 2008), pp 3213-3218, ISSN 0278-6915

Weissinger, J. (1994). *Animal Drugs and Human Health (*fourth edition*)*, L.M. Crawford, and D.A. Franco, Technomic Publishing Co, ISBN: 9781566761024, Lancaster, USA

World Health Organization, WHO. (1989). In Evaluation of Certain Veterinary Drug Residues in Food, Thirty-fourth Report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series 788

World Health Organization,WHO.(1993). In Evaluation of Certain Veterinary Drug Residues in Food, Fortieth Report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series 832

Zinedine, A.; Soriano, JM.; Moltó JC. & Mañes J. (2007). Review on the toxicity, occurrence, metabolism, detoxification, regulations and intake of zearalenone: an oestrogenic mycotoxin. *Food and Chemical Toxicology*,Vol. 45, No. 1, (January, 2007), pp 1-18 ISSN 0278-6915

9

Viable but Nonculturable Bacteria in Food

Marco Sebastiano Nicolò and Salvatore Pietro Paolo Guglielmino
University of Messina
Italy

1. Introduction

> "Dis-moi ce que tu manges, je te dirai ce que tu es." (Tell me what you eat and I will tell you what you are - Anthelme Brillat-Savarin, *Physiologie du Goût, ou Méditations de Gastronomie Transcendante,* 1826)
> "Ninety per cent of the diseases known to man are caused by cheap foodstuffs. You are what you eat." (Victor Lindlahr, 1923)

In every time, availability of food has been a struggle for human survival. In particular, food storage techniques and maintaining have represented one of the most important milestones in the evolution and development of human society. In that sense, the use of fire for cooking and salt and spices for conservation have been important discoveries in food processing, with immediate consequences on human habits and life.

Nowadays, food resources remain a primary objective for human society, specifically in terms of safety and control.

In fact, up to date more than 250 foodborne diseases have been described.

Infectious foodborne diseases are caused by the consumption of food contaminated by pathogen bacteria, viruses, parasites and prions.

Contamination may occur as consequence of incorrect practices at different steps of food processing, such as handling of feedstock, decontamination, packaging and storage.

The rapid globalization and trade among countries with different hygienic standards have increased the possibility of food contamination. Hence, outbreaks of foodborne diseases that were once contained within a small community may now take place on global dimensions.

The fundamental strategy for outbreaks monitoring is the traceability, that is the possibility to identify the pathogen(s), the ways of food contamination and the spreading.

One of the most critical problems in traceability of bacterial pathogens is represented by a state of latency of most foodborne bacterial species, called "viable but nonculturable (VBNC) state", induced by environmental stresses, such as low temperature, high osmolarity, and nutrient starvation.

In such a state, bacteria show a discrete metabolic activity, but are not able to replicate.

Following environmental stimuli, as temperature shift or replenishment of nutrients, VBNC bacteria can "resuscitate", so restoring their ability to grow on common culture media.

Still open is the debate about the possibility, for pathogen bacteria in VBNC state, to maintain pathogenicity and trigger disease in their hosts.

The evidence of contrasting results, in fact, indicates the need for a better understanding of such complex phenomenon, particularly about the underlying molecular network.

Several studies have shown that the most of human pathogens, especially foodborne bacteria, may enter VBNC state as survival strategy against environmental stress.

Many chemo-physical characteristics in food (as acidic pH, low content in carbohydrates, etc.), as well as processing, decontamination and storage, may induce VBNC state.

The observation of VBNC forms of foodborne bacteria and the lack of a ultimate information about the possibility for VBNC bacteria to retain their virulence has raised the problem about the necessity of new procedures for VBNC detection in food.

Many systems have been proposed for VBNC detection in water, but their application on food seems to be quite difficult. In fact, factors as food texture and pH, as well as presence of free lytic enzymes and other compounds, may interfere with the chemical reaction(s) required for the assay.

Then, traceability of foodborne VBNC pathogens strongly requires the design of new detecting systems.

2. The causes of infectious foodborne disease outbreaks

Infectious foodborne diseases are caused by the consumption of food or beverages contaminated by many pathogenic bacteria, viruses, parasites and prions. In addition, contamination can also be due to molecules produced by bacteria, called **toxins**. Toxins can be derived from cell structure, such as lipopolysaccharide (LPS) of Gram negative bacteria (**endotoxins**), as well as they are synthesized inside the cell and secreted in the surrounding environment (**exotoxins**), as botulinum, cholera and Shiga toxins.

After food contamination, the microbe or its toxin enters the organism via the digestive tract, triggering the illness.

In general, an outbreak of foodborne disease occurs when a group of people eats the same contaminated food and two or more of them develop the same illness.

Many outbreaks consist of sporadic cases and are self-limiting, in that they are related to a small quantity of contaminated food which, usually, is totally eaten by few people and involve a specific geographic area. A sporadic outbreak may follow a catered meal or eating a meal at a restaurant on a particularly busy day.

The number of people affected and the extension of geographic areas in which a foodborne outbreak occurs have considerably increased as consequence of the variations in social habits. Commonly, workers and students may have meals at large-scale retail trade structures as canteens, highway stops, fast food chains and refectories. Such structures, usually, look to suppliers which, from the place where food is produced, distribute their products on national scale. Then, contaminated food may be transported in different places, causing many distinct outbreaks at the same time within a country.

For example, in 2005, in South Wales an outbreak of *Escherichia coli* O157 occurred, with identification of 157 cases, 31 people hospitalized, and one 5-year old child died, as well as in 2008, an outbreak of listeriosis in Canada (57 clinical cases) killed 23 people.

Recently, in October 2010 an outbreak of cholera occurred in Haiti. On 18 August 2011, the total number of reported cholera cases was 419,511, with 222,359 hospitalized. Overall, data from health facilities indicate that 5,968 people have died (case fatality rate 1.4%).

Moreover, actually it has been recognized that outbreaks may spread on wider geographic areas. In fact, the rapid globalization and trade among countries with different hygienic standards has increased the risk and entity of food contamination. As consequence, many outbreaks that were once contained within a small community may now take place on wider geographic areas and several examples can be considered.

The Bovine spongiform encephalopathy (BSE), commonly known as **mad-cow disease**, was firstly detected in United Kingdom and represented a world-wide crisis. The cause of global infection was attributed to the use of contaminated bone meal, used in livestock feeding. Only in the United Kingdom, 4.4 millions of cattle were slaughtered in the attempt to eradicate the pathology and limit its spreading; moreover, the European Union banned exports of British beef from March 1996 to May 2006.

In 2011, an outbreak of gastroenteritis caused by *E. coli* O104:H4 in Germany had several social, political and economic implications throughout whole Europe.

Often, such kind of outbreaks are very difficult to identify while occurring. An example is represented by an outbreak of *Salmonella* which simultaneously happened in Europe, North America and Israel, following the importation of contaminated snack food. The outbreak identification was accidental, because it was caused by a rare strain of *Salmonella* and the number of cases in each geographical entity was not high.

Therefore, there is a strong probability that contaminated foodstuffs distributed in different countries could origin outbreaks which are not readily pointed out by national health authorities if the spreading in each single country is limited. Many sporadic cases happening in single countries may be, then, part of a common global outbreak.

3. The causes of food contamination

Contamination may occur as consequence of incorrect practices at different steps of food processing, such as feedstock production and handling, food preparation, packaging and storage. Each stage of food processing shows critical points.

3.1 Feedstock production

Use of raw manure or sewage for fertilization and lack of hygiene and health controls in fish and cattle breeding are the first cause of contamination. Spinach and lettuce have been linked to *E. coli* O157:H7 outbreaks (Erickson et al., 2010). Untreated food, such as fresh fruit juice and milk, carry risks because they are not subjected to any decontaminating procedure.

Similarly, incorrect slaughterhouse practices can lead to contamination, especially when faecal or intestinal matter from cattle mixes with the meat.

3.2 Feedstock handling

Handling and washing of feedstock before processing are very important sources of contamination. Handling by ill people which have cuts, open sores or skin infections causes spreading of *Staphylococcus aureus*, which is often found on skin with boils and blisters. That is why food handling should be performed by personnel wearing gloves. However, touching contaminated surfaces, coughing into a gloved hand or handling money before food preparation can still spread germs, which is why gloves should be changed often.

3.3 Food preparation

Beyond the causes of contamination above discussed, in this step cross contamination is a crucial factor. This can happen during food preparation and storage. For example, juices from raw beef and poultry could mix with ready-to-eat food or when kitchen tools are used without distinction for both raw beef and fresh vegetables. Cleaning tools with soap and hot water strongly reduces contamination.

Also undercooking can be cause of food contamination, because heat drastically reduces the presence of bacteria in food.

3.4 Food packaging

In food packaging, hygiene of handlers and materials employed are critical parameters.

Historically, canned food, if prepared under unsafe conditions, can be susceptible to contamination by toxinogenic microbes, such as *Clostridium botulinum*.

Nowadays, food packaging is mostly based on plastics, employed for vegetables, cheese, fruit. However, even if a correct packaging assures food safety from microbe contamination and remarkably prolongs shelf-life, much attention has to be paid about hygienic requirements. Bacteria can colonize and adhere on plastics, producing biofilms which may reduce shelf life or, for pathogenic bacteria, they may secrete toxins or proliferate on food.

3.5 Food storage

This is a critical step for perishable food which requires refrigeration or freezing, because low temperature slows down bacterial reproduction, even if some species, called **psycrophiles**, such as *Listeria monocytogenes*, *Yersinia enterocolitica* and *Pseudomonas* sp., are able to develop also at refrigerator temperature.

Food storage has become increasingly important in the perspective of globalization, in that different food tipologies may be transported together. Physical contact and dripping of juices to other foodstuffs are causes of crosscontamination. In this way, one kind of food can be contaminated by unusual bacteria, that become difficult to be identified. Also changes in livestock farming and industrialization of slaughtering of pigs have played an important role. Transportation of live animals for slaughtering has in some studies been proved to be an important factor in dissemination of *Y. enterocolitica* from farm to farm (Nesbakken, 2007).

4. Foodborne bacteria

Bacteria are single-cell microrganisms colonizing any environment. From an ecological point of view, they play a pivotal role in the biogeochemical cycles, by which chemical elements move from living to non-living matter and vice versa. Some of them produce molecules which improve the quality of human life, such as antibiotics, vitamins, probiotics and other nutritional factors.

In food industry, several *Lactobacillus* species are very important for production of cheese, yogurt, beer, cider, wine, bread and chocolate, as well as in functional food.

Several bacteria inhabit the skin and the intestine of animals and humans, protecting them from contamination by hazardous microbes and playing a role in metabolism.

Pathogen bacteria are responsible for infectious diseases, by colonization of tissues and organs, whose physiology is altered by bacterial metabolism and reproduction.

Foodborne pathogen bacteria are a group of microrganisms which can contaminate food and, after swallowing, cause illness.

Common symptoms of foodborne illness are diarrhoea and/or vomiting, abdominal cramps, nausea, fever, joint/back aches, and fatigue.

A description of the most common known foodborne pathogen bacteria and related illness(es) follows. Intriguingly, for the most of them, the VBNC state has been described.

Bacillus cereus

Gram positive rod-shaped bacterium, able to form endospores under negative conditions, commonly found in soil and vegetation. After pasteurisation or heating, endospores survive, whereas competing microflora is eliminated. During food cooling, endospores germinate and vegetative cells proliferate, producing several toxins, one of which is highly resistant to heat and to pH between 2 and 11. The infection may be almost diarrhoeal, but some cases described nausea and vomiting. The illness is self-limiting.

Brucella spp.

Genus of gram negative, aerobic, rod-spherical shaped bacteria which infect animals and humans. Four species are recognised as harmful, that is *B. abortus* (from cattle), *B. melitensis* (from goats, sheep, and camels), *B. suis* (from pigs), and *B. canis* (from dogs).

It can contaminate raw milk and cheese, but may be acquired also by inhalation, causing **brucellosis**, also known as Malta fever, undulant fever, Rock of Gibraltar fever, and Bang's disease.

The symptoms are non-specific and systemic, with fever, sweats, headache, anorexia, back pain. The chronic form causes suppurative lesions in the liver, spleen, and bone, with a mortality of 5% in untreated individuals.

Campylobacter jejuni

Gram negative, microaerophilic, spirally curved bacterium, commonly living in the bowel of animals, especially in poultry, and spread by faeces and milk.

It can also contaminate incorrectly prepared meat and poultry.

It is one of the first cause of foodborne illness (**campylobacteriosis**) in United States and in United Kingdom. Symptoms of campylobacteriosis are fever, cramping abdominal pain and diarrhoea. Remission follows within a week.

Clostridium botulinum

Gram positive rod-shaped bacterium, obligate anaerobe, which forms endospores in presence of oxygen or other environmental stresses. Food contamination can occur due to improperly preserved or home-canned, low-acid food, which allows endospore germination and subsequent toxin secretion. Among the toxins, seven are neurotoxins, that cause the **flaccid muscular paralysis** by inhibition of neuromuscular transmission through decreased acetylcholine release. Death occurs when respiratory mechanics is compromised.

Clostridium perfringens

Gram positive rod-shaped bacterium, obligate anaerobe, which forms endospores in presence of oxygen or other environmental stresses. Preferentially, it contaminates meat-based food, because aminoacid content satisfies its nutritional requirements. After cooking, oxygen concentration is lowered and endospores germinate during cooling. Storage time and refrigeration are critical, in that the vegetative cell can duplicate in 20 minutes. After ingestion, in the bowel the bacterium produces an exotoxin which causes abdominal pain and diarrhoea. In the most of cases, the illness is self-limiting.

Corynebacterium ulcerans

Gram-positive, nonmotile, straight to slightly curved rod-shaped bacterium that causes subacute bovine mastitis, but *C. ulcerans* has increasingly been isolated from domestic animals such as dogs and cats.

Consumption of raw milk and dairy products or contact with cattle may cause infection of humans, causing a disease very similar to diphtheria by secretion of a toxin which inhibits protein synthesis of epithelial cells, leading them to death.

Symptoms are sore throat, low fever, and a pseudomembrane on the upper respiratory tract.

Toxin may spread through the bloodstream and can lead to life-threatening complications in heart and kidneys. It can also cause nerve damage, eventually leading to paralysis. 40% to 50% of those left untreated can die.

Coxiella burnetii

C. burnetii is an obligate intracellular, small Gram-negative bacterium highly resistant to high temperature, osmotic pressure, ultraviolet light.

It can contaminate people after ingestion of raw milk or contact with infected animals.

C. burnetii is highly infectious via the respiratory route, but the infectivity via the oral route is poorly understood. It has been proposed that *C. burnetii* can escape the gastrointestinal tract and produce infection sufficient to stimulate systemic immunity (Loftis et al., 2010).

C. burnetii secretes a toxin inside the phagolysosome which inhibits its fusion with the cell degradation endosomes.

In humans it causes the **Q fever**, with fever, severe headache, muscle and joint pains, upper respiratory problems and gastro-intestinal symptoms such as nausea, vomiting and diarrhoea. Life threatening complications are acute respiratory distress syndrome (ARDS), granulomatous hepatitis and retinal vasculitis.

The chronic form of Q fever is virtually identical to endocarditis. It is usually fatal if untreated; however, with appropriate treatment the mortality falls to around 10%.

Escherichia coli

Gram negative, rod-shaped bacterium, inhabitant of mammal bowel. Among the pathogenic strains, distinct groups can be identified.

- Enterotoxigenic (ETEC) *E. coli* produces and releases two exotoxins, LT (heat-labile) enterotoxin, similar to cholera toxin, and ST enterotoxin, which causes cGMP accumulation in the target cells and a subsequent secretion of fluid and electrolytes into the intestinal lumen. Both toxins induce watery diarrhoea, similarly to cholera. It is responsible of the majority of "traveller's diarrhoea" and infant diarrhoea in developing countries.
- Enteroinvasive (EIEC) *E. coli* can invade the intestinal wall, giving inflammation, fever and diarrhoea similar to *Shigella*-like dysentery.
- Among the enterohaemorrhagic (EHEC) *E. coli* strains, the most important is the well known serotype O157:H7. EHEC strains produce Shiga-like toxins, which induce haemorrhagic colitis, resulting in bloody diarrhoea. Sometimes the toxin may spread in kidney, causing a very dangerous complication called **haemolytic uraemic syndrome**. The first symptom is the presence of blood in urine leading to kidney failure.
- Enteropathogenic (EPEC) strains use an adhesin known as **intimin** to bind host intestinal cells. Adherence to the intestinal mucosa causes a rearrangement of actin in the host cell, causing significant deformation. EPEC cells are moderately invasive and elicit an inflammatory response. Changes in intestinal cell ultrastructure are likely the prime cause of diarrhoea in those afflicted with EPEC.
- Enteroaggregative (EAEC) strains have fimbriae which aggregate tissue culture cells. EAEC strains bind to the intestinal mucosa to cause watery diarrhoea without fever and are not invasive. They produce a haemolysin and a ST enterotoxin similar to that of ETEC strains.

Helicobacter pylori

Gram negative, spiral-shaped microaerophilic bacterium, normally colonising the stomach of 30-50% of the human population in developed countries.

The ways of transmission are likely to be oral and oro-faecal routes, even if it has been quite difficult its isolation by the faeces (Thomas et al., 1992; Kelly et al., 1994).

In the stomach, the bacterium secretes the enzyme urease, which degrades urea to ammonia. Ammonia lowers the pH locally, so permitting the development of the infection.

It is responsible of chronic gastritis, peptic ulcer disease and stomach cancer. Actually, it is also associated with the gastric MALT lymphoma and its role on other several illnesses, as Sjögren, Prader-Willy and Raynaud syndromes, is matter of discussion.

Listeria monocytogenes

Gram positive, facultative anaerobe, rod-shaped intracellular bacterium, commonly spread in soil and water. Vegetables are contaminated by the soil or by manure used as fertilizer. Animals can carry the bacterium asymptomatically and can contaminate food, such as uncooked meat, raw milk and soft cheeses, that are usually prohibited to pregnant women.

It is the causative agent of **listeriosis**. After having entered immune system cells, it becomes septicemic and can grow. The intracellular state in phagocytic cells may permit access to the brain and probably to reach the fetus in pregnant women by crossing the placenta.

Listeriosis may arise as septicaemia, meningoencephalitis, corneal ulcer, pneumonia and during pregnancy (causing abortion within 6-9 months or stillbirth).

When no internalization in cells occurs, the disease is presented as a febrile gastroenteritis.

L. monocytogenes is the leading cause of death among foodborne bacterial pathogens, with 20 to 30 percent of clinical infections resulting in death.

Plesiomonas shigelloides

Gram-negative, rod-shaped bacterium, which has been isolated from freshwater, freshwater fish, and shellfish and from many types of animals including cattle in tropical and sub-tropical areas.

Common symptoms are fever, shivers, abdominal pain, nausea, diarrhoea (which, in severe cases, may be greenish-yellow, foamy, and bloody) or vomiting and appear 20-24 hours after consumption of contaminated food or water.

P. shigelloides gastroenteritis is usually a mild self-limiting disease with remission in healthy people within 1-7 days, described in African countries, but sporadic cases have been reported also in Europe and North America. Severe forms may include sepsis and meningoencephalitis in newborns, arthritis, cholecystitis and osteomyelitis (Terpeluk, 1992).

***Salmonella* spp.**

The genus *Salmonella* consists of Gram negative, rod-shaped facultatively anaerobe bacteria, which live in the bowel of humans and animals.

Meat, poultry and eggs are the most common food contaminated by *Salmonella*. However, food is contaminated by low loads of *Salmonella*, which are quite difficult to be appreciated by standard detection methods.

All species are considered pathogenic and responsible for a gastroenteritis known as **salmonellosis**. Bacteria reach the bowel and duplicate. Sometimes, they may cross the mucosa, entering lymphatic and cardiovascular systems; from there, they may infect other organs, causing septicaemia.

Symptoms are moderate fever, nausea, abdominal pain and cramps, diarrhoea. Average mortality rate is below 1%, but increases in children and old people, occurring as septicaemia.

Among the species, *S. typhi* is the most virulent. It is the aetiological agent of **typhoid fever** and is spread only by human faeces. Symptoms are high fever (40°C) and continued

headache, then diarrhoea appears and fever declines. Differently from salmonellosis, bacteria multiply into phagocytic cells and disseminate in the body. In severe cases, perforation of bowel mucosa may occur.

Today, the mortality rate of typhoid fever is about 1-2%; in the past it exceeded also 10%.

***Shigella* spp.**

The genus *Shigella* is composed of Gram negative, rod-shaped bacteria, which are normally present in the bowel of humans, apes and monkeys.

Food contamination occurs via handling by unhealthy operators. Four species are pathogenic: *S. sonnei*, *S. dysenteriae*, *S. flexneri* and *S. boydii*. Many cases of "traveller's diarrhoea" are supported by *Shigella* spp.

S. dysenteriae causes severe dysentery and prostration by secreting the "Shiga toxin", which inhibits protein synthesis in epithelial cells of intestinal wall, killing them. Bacteria multiply in the small intestine and spread in the large intestin, entering the epithelial cells. Infection proceeds towards neighbouring cells, via a "cell-to-cell" mechanism, thus avoiding immune system response. As consequence, intestinal mucosa is damaged, causing severe diarrhoea with blood and mucus in the faeces. Additional symptoms may be abdominal cramps and fever.

Only *S. dysenteriae* may reach the bloodstream, causing septicaemia. Its mortality rate is quite significant and may reach 20% in tropical areas, where it is prevalent.

Staphylococcus aureus

Gram positive, spherical-shaped bacterium, it is a frequent inhabitant of respiratory tract from which it can contaminate hands and skin. Typically, *S. aureus* and others staphylococcal species are resistant to various stresses. They can survive 60°C for 30 minutes, drying and high osmotic pressures. Such characteristics allow their survival under conditions that, generally, eliminate the most of bacteria. In absence of competitors, then, it can rapidly proliferate.

It is able to produce several toxins that improve the virulence or damage tissues, but the toxin produced by serogroup A is responsible for most of the cases. The toxin is heat-stable, in that maintains its virulence up to 30 minutes of boiling.

Symptoms of intoxication are vomiting and abdominal cramps followed by diarrhoea. Remission is reached within 24 hours.

***Streptococcus* spp.**

Genus of gram positive, microaerophilic, spherical-shaped bacteria which occur in chains or pairs. The genus is defined by a combination of antigenic, haemolytic, and physiological characteristics into Groups A, B, C, D, F, and G.

Groups A and D can be transmitted to humans via food.

Group A is formed by a single species, *S. pyogenes*, with 40 antigenic types, while Group D is represented by the new genus *Enterococcus*.

Many types of food may be contaminated by Group A *Streptococcus*, including raw milk, ice cream, ham, potato salad, shrimps salad, steamed lobster via manipulation by ill handlers. Moreover, they may contaminate food with low content of free water, rich in salt or with very acidic pH, which is generally difficult to be contaminated by other bacterial species.

Group A *Streptococcus* may cause several diseases, as bacteraemia, impetigo, erysipelas, pneumonia, osteomyelitis, septic arthritis, meningitis and toxic shock syndrome.

Among complications, acute rheumatic fever may follow respiratory infections as an autoimmune disease, in which antibodies raised against the streptococcal M-protein cross-react with autoantigens of pericardium and synovium.

Enterococci can cause food intoxication through production of biogenic amines, such as histamine and tyramine, mainly by the decarboxylation of amino acids in fish, meat and dairy products, wine, beer, vegetables, fruits, and nuts.

The genus *Enterococcus* is responsible for severe diseases, such as urinary tract infections, bacteraemia, bacterial endocarditis, diverticulitis, and meningitis.

Vibrio cholerae

Gram negative, slightly curved rod-shaped bacterium, which lives naturally in brackish waters, but can easily spread also in freshwater.

It can contaminate fish and seafood.

It is the aetiological agent of cholera, a serious illness that stroke Europe and North America during the XIX century with different outbreaks. Symptoms are nausea, vomiting, abdominal pain, watery diarrhoea which may induce severe dehydratation, that can be fatal if untreated.

Nowadays, cholera is endemic in India and rarely it causes outbreaks in Western countries. In 1991-1994, Latin America had an epidemic, in consequence of importation of contaminated seafood from Asia, with more than one million cases and 9600 deaths. In 2010, a new outbreak in Haiti has occurred.

Vibrio parahaemolyticus

Gram negative, slightly curved rod-shaped bacterium, which lives naturally in salt waters.

The related gastroenteritis follows the consumption of contaminated seafood. Symptoms include abdominal pain, vomiting, a burning sensation in the stomach and cholera-like watery faeces. Remission occurs within few days.

Vibrio vulnificus

As *V. parahemolyticus*, it is found in estuarine waters and contaminates seafood. It represents a serious threat for people with compromised immune system. In people with liver disease, it may cause septicaemia, with mortality rates often exceeding 50%.

Yersinia enterocolitica* and *Y. pseudotuberculosis

Psycrofile Gram-negative bacteria which usually colonise the bowel of domestic animals and are transmitted by milk and meat. Outside their natural habitat, they are able to grow at refrigeration temperature (4°C). The symptoms are diarrhoea, fever, headache and abdominal pain.

For the most of the above described foodborne bacteria, the VBNC state has been proven by several studies. Therefore, the meaning of VBNC state as well as inducing and resuscitating factors have to be well understood, in order to plan how to face the difficulties of detection and make traceability feasible.

5. The bacterial stress response

In their natural environments, bacteria undergo fluctuating chemo-physical conditions, such as nutrient availability, temperature, osmolarity, and pH, which may interfere with their growth and survival. In such a situation, they modulate their gene expression, in order to survive, by activating the **bacterial stress response**, which consists in variations of cell morphology, dimensions, energetic levels, directly related to cell survival. The final effect is an increased global resistance of surviving cells against further stresses, such as exposure to antibiotics, hydrogen peroxide and high osmolarity (Matin, 1991; Nyström et al., 1992).

In such a situation, some genera of Gram positive bacteria, as *Bacillus*, *Clostridium* and few others, differentiate into **endospore**.

Endospore is a differentiated bacterial cell in which new structures, required for mechanical and physical resistance, are synthesised. In the endospore, metabolism has been abolished by a controlled process of dehydration. The lack of metabolism confers a global resistance extended for prolonged periods of time. When environmental conditions become favourable, endospore undergoes **germination**, a process by which the metabolic activity is restored and cell turns to vegetative life.

The most of Gram positive and all Gram negative bacteria are not able to generate endospores. However, they can trigger a stress response with a highly-complex network of molecular mechanisms.

Studies carried on bacterial stress response have started from bacterial physiology in natural oligotrophic environments (Kurath & Morita, 1983; Morita, 1982).

Subsequent studies have focused on the underlying molecular mechanisms, especially regarding a specific chemophysical stress and single-nutrients starvation (Eberl et al, 1996; Matin, 1991; Nakashima et al., 1996; Rockabrand et al., 1995).

One aspect on which little is known is that, in natural environments, nutritional and chemo-physical factors inducing bacterial stress response may change simultaneously, with one specific factor influencing another and being, in turn, influenced by a third one.

Therefore, several mechanisms of bacterial stress responses should be activated. Unfortunately, few data are available about bacterial response to simultaneously-acting multiple stresses.

5.1 The bacterial stress response to chemophysical stress

Bacterial responses to chemophysical stresses, such as cold- and heat-shock, hyperosmolarity and acid pH, have been investigated.

5.1.1 The cold shock response

It has been observed that *E. coli*, when subjected to a temperature decrease, triggers the cold shock response, in which sets of proteins are synthesised, globally indicated as CIPs (Cold

Induced Proteins), as transcription regulators, ribosomal proteins, elongation factors and β subunit of RNA-polymerase (Berger et al., 1996; Jones & Inouye, 1996; Nakashima et al., 1996), in order to allow mRNAs transcription and protein synthesis.

5.1.2 The heat shock response

When subjected to heat shock, bacteria have to counteract macromolecules denaturation, so the overexpression of chaperones is induced to stabilize cell macromolecules and to degrade denaturised polypeptides (Gage & Neidhardt, 1993; Spence, 1990).

5.1.3 The osmotic shock response

Osmotic shock causes considerable shrinkage of the cytoplasmic volume. In this process, known as **plasmolysis**, intracellular water tends to migrate towards the external environment. As a consequence, the concentrations of all the intracellular metabolites increase and thus reduce the intracellular water activity (w_a). Bacteria may react by increasing the concentrations of solutes that have no effects on cell processes. Such solutes, called **compatible solutes**, are potassium ions, some amino acids (such as glutamate, glutamine, proline, γ-aminobutyrate, alanine), the quaternary amines glycinebetaine and some sugars (sucrose, threalose). Their accumulation inside the cells allows the cytoplasmic w_a to be restored, with maintaining of metabolism (Csonka, 1989).

5.1.4 The acid shock response

When a sudden drop in pH occurs, bacteria generally use proton pumps, which literally pump protons out of the cell to keep the cytoplasmic pH. Another approach is to increase the concentration of alkaline compounds within the cell to counteract the acidification of the cytoplasm (Bore et al., 2007).

5.2 The bacterial stress response to nutrient starvation

Under nutrient starvation, variations at structural, metabolic and physiological levels are observed.

During carbon starvation, heterotrophic bacteria enter a quiescent state, with a decrease in ATP content and a general reduction in metabolic activity. Moreover, gram negative rod bacteria undergo a characteristic morphology transition to spherical shape. When an energetic source becomes newly available, carbon-starved bacteria rapidly resume their metabolic activities, simultaneously restoring the rod shape (Givskov et al., 1994).

On the other hand, under nitrogen or phosphate starvation, anabolism is strongly reduced and the cell undergoes an energetic surplus, which is dissipated by futile cycles reactions or accumulation of high-energy storage compounds, such as PHAs (Eberl et al., 1996).

6. The viable but nonculturable (VBNC) state

One of the most intriguing findings on stressed bacteria in natural environments was the so-called **viable but nonculturable** (VBNC) **state** (Xu et al., 1982).

The VBNC state is defined as a state of dormancy triggered by environmental harsh conditions, such as nutrient starvation (Cook & Bolster, 2007), temperature (Besnard et al.,

2002), osmotic stress (Asakura et al., 2008), oxygen availability (Kana et al., 2008), several food preservatives (Quirós et al., 2009), heavy metals (Ghezzi & Steck, 1999), exposure to white light (Gourmelon et al., 1994) and decontaminating processes, as pasteurization of milk (Gunasekera et al., 2002) and chlorination of wastewater (Oliver et al., 2005).

In such a state, bacteria lose the ability to grow on solid media and undergo reduction in size; moreover, several metabolic variations occur, such as reductions in nutrient transport across cytoplasmic membrane, respiration rates, and macromolecular synthesis (Oliver, 2000; Porter et al., 1995). Biosynthesis does not cease, in that starvation and cold shock proteins are synthesized (McGovern & Oliver, 1995; Morton & Oliver, 1994). ATP levels remain high in VBNC cells (Beumer et al., 1992; Federighi et al., 1998). Further, recent studies have demonstrated continued gene expression by cells in the VBNC state (Lleò et al., 2000, 2001; Yaron and Matthews, 2002). Other cellular characteristics, such as cell wall (Signoretto et al., 2000; Signoretto et al., 2002) and membrane composition (Day and Oliver, 2004), differ remarkably from culturable cells.

When environmental conditions become permissive, **resuscitation** of VBNC bacteria occurs. Resuscitated bacteria are culturable on solid media and display the vegetative lifecycle.

Also resuscitation is a very complex phenomenon. A group of extracellular proteins, indicated as **resuscitation promoting factors** (Rpfs), play a key role in several other bacterial species (Hett et al., 2007; Mukamolova et al., 1998a, 1998b; Shleeva et al., 2004).

Another class of resuscitation factors is a heat-stable **autoinducer of growth** (Reissbrodt et al., 2002), which has been identified as a novel quorum-sensing system, termed AI-3 (Sperandio et al., 2003) and secreted after incubation in media containing norepinephrine (Freestone et al., 1999). Norepinephrine is produced in large amounts in humans following severe tissue injury, and is thus considered to be a stress-related hormone. Both epinephrine and norepinephrine could replace AI-3 in activating enterohaemorrhagic *E. coli* virulence gene expression (Sperandio et al., 2003).

These findings would support the hypothesis of resuscitation of VBNC enteropathogens in the human intestinal tract, at a time (e.g. tissue damage) when the host may be under significant physiological stress, with consequent secretion of norepinephrine.

Interestingly, even several higher organisms may induce resuscitation from the VBNC state.

Many conditions have been found to allow resuscitation of pathogens, as inoculation into yolk sacs of embryonated eggs (Cappelier et al., 1999b, 2007), into mice (Cappelier et al., 1999a) and into human volunteers (Colwell et al., 1996).

The observations of the VBNC state for the most of foodborne pathogens have raised several questions about the retention of virulence in such a state as well as recovery of virulence together with resuscitation in the host. The matter is highly debated because contrasting results have been presented. In some cases, virulence of *L. monocytogenes* in the VBNC state has been shown to depend on the experimental conditions adopted for resuscitation (Cappelier et al., 2005, 2007). It seems that VBNC pathogens are not generally able to initiate disease, but virulence is retained and infection can be initiated following their resuscitation. In fact, VBNC cells of *Vibrio harveyi* were avirulent, but resuscitated cells were lethal, indicating that VBNC *V. harveyi* cells retained pathogenic potential (Sun et al., 2008). Similarly, Oliver & Bockian (1995) reported *V. vulnificus* to lose virulence for mice in

proportion to the length of time that the cells were in the VBNC state. The cells retained virulence, however, and even when fully nonculturable, were able to cause fatal infections, with resuscitation occurring within the mouse. Continued virulence for a variety of pathogenic vibrios has also been demonstrated (Baffone et al., 2003).

Then, it can be concluded that VBNC forms of foodborne pathogen bacteria into a human or animal host may be a realistic threat for public health, because it has been demonstrated that virulence can be maintained or recovered after resuscitation.

7. Role of food in induction of VBNC state

Chemophysical characteristics of food select the bacteria able to colonize and survive.

Such characteristics, defined as **intrinsic factors**, are pH, redox potential (or oxygen concentration) and a_w. However, other factors, that have to be considered, will be also discussed.

7.1 pH

Although optimal bacterial growth requires a pH near neutrality, it has been demonstrated that foodborne pathogens can trigger the acid shock response for survival in acidic environments.

Then, acidic food may be an environment inducing the VBNC state and risk becomes higher for food, as fresh juices and salads, not subjected to antimicrobial treatments.

7.2 Redox potential (or oxygen concentration)

Food redox potential is a factor which influences the development of bacteria. Vegetable food has a redox potential ranging from +300 and -400 mV, which favours aerobes and eukaryotic microrganisms (yeasts and moulds). Meat has a value of -200 mV, which allows persistence of microaerobes and anaerobes bacteria.

Oxygen influences food redox potential, so its concentration is critical for bacterial growth. Aerobe bacteria need oxygen for growth, whereas anaerobes use compounds other than oxygen, such as nitrate, nitrite, sulphate, sulphite, etc. Some anaerobes, defined **obligate**, have been considered as totally inhibited by presence of oxygen for a long time, in that were believed to be unable to face the oxidative damage caused by Reactive Oxygen Species (ROS). Nowadays, it has been demonstrated that *Clostridium* may withstand oxidative stress, in that genome sequencing revealed the presence of genes related to oxygen metabolism (Kawasaki et al., 2005).

Also oxidative stress may induce VBNC state, as demonstrated in *Vibrio* sp. (McDougald & Kjelleberg, 2006).

7.3 Water activity (a_w)

The a_w is a measure of free water available for microbial growth and its value ranges from 1.00 of mineral water to 0.60 of bakery products, candies and dried fruit. A range of a_w around 0.995 to 0.998 allows the growth of most of bacteria. The use of salting as ancient

conservation technique for meat and vegetables consists, ultimately, in bringing down a_w to reduce microbial development.

As already stated, bacteria may face osmotic shock; moreover, high osmolarity may induce VBNC state (Asakura et al., 2008).

7.4 Chemical composition

Biological macromolecules are used as substrates for growth, but bacteria may use only a limited part of the macromolecular content. Moreover, several compounds, such as polyphenols of fruit and vegetables and organic acids, display bacteriostatic and bactericide activities, which are further stress-inducing factors.

7.5 Feedstock treatment

Decontamination procedures, such as pasteurisation, salting and acidification, can induce VBNC state (Gunasekera et al., 2002; Makino et al., 2000; Quirós et al., 2009). Moreover, untreated food is more susceptible to microbial contamination (Erickson et al., 2010).

7.6 Food storage

Before consumption, food is stored during packaging, distribution and commercialization. Time and temperature are critical factors which are related to bacterial stress response and induction of VBNC state (Besnard et al., 2002).

7.7 Presence of endogenous microflora

Feedstock and untreated food possess a resident microflora, not necessarily hazardous for human health. Such microflora may interact with contaminating bacterial pathogens by intercellular exchange of signalling molecules, as the **autoinducers**, which may activate secretion of virulence factors or resuscitation in VBNC bacteria (Reissbrodt et al., 2002).

8. VBNC state of foodborne bacteria – The new challenge in food safety

Despite the few studies, strong evidences of VBNC bacteria in food have been reported. In stored wine, for example, acetic acid and lactic acid bacteria entered VBNC state as consequence of lack of oxygen and presence of sulphites, respectively (Millet and Lonvaud-Funel, 2000).

The role played by chemo-physical characteristics of food in triggering VBNC state in foodborne pathogen bacteria is poorly undestood. The most of the studies consists in analysing the effects provoked by a single stress on a homogeneous population of bacteria.

However, it has to be considered that, in their habitat, bacteria are continuously subjected to the simultaneous action of factors, as nutrient availability, pH, osmolarity, temperature, presence of toxic compounds, ecological competition with other organisms, which are continuously changing. The time and the entity of the change of a single factor influences, and in turn is influenced by, changes and entities of other factors, according to chaos dynamics observed in nature, in an unpredictable way.

In the same way, food and its surrounding environment have to be considered as a complex system, in which the chemo-physical characteristics (pH, a_w, chemical composition) and environmental factors (storage temperature and time, decontamination treatments, packaging under modified atmosphere) act simultaneously on contaminating bacteria.

It has been demonstrated (Nicolò et al., 2011) that refrigerated pasteurised grapefruit juice induces VBNC state in *E. coli* O157:H7 and *S.* Typhimurium within 24 hours of incubation.

Grapefruit has a very acidic pH, low content in carbohydrates and several antimicrobial compounds. Such characteristics, generally described as factors inducing VBNC state in laboratory, together with the refrigeration used for storage, suggested the hypothesis that grapefruit juice could induce VBNC state in foodborne pathogens.

On the contrary, grape juice, which differs from grapefruit juice for the higher content in carbohydrates, did not induce VBNC state (Nicolò et al., unpublished data), despite the acidic pH and refrigeration temperature.

Therefore, the role of food in induction of VBNC state has to be elucidated. Predictive models offered by biomathematics and bioinformatics would be very helpful tools, in order to evaluate the possibility that, under certain conditions, pathogen bacteria contaminating a tipology of food may enter the VBNC state.

In pasteurised milk, de novo expression of a *gfp* reporter gene has been demonstrated to be higher than culturable cells for both *E. coli* and *Pseudomonas putida*, so showing that, after thermal treatment, contaminant bacteria had lost the ability to form colonies, but retained transcription and translation machineries (Gunasekera et al., 2002).

A study on dried salted squid contaminated by *Salmonella enterica* subsp. *enterica* Oranienburg, responsible for an outbreak in Japan in 1999, has showed the inefficacy of cultural methods for detection of VBNC bacteria. In such a study, less than 20 culturable cells were recovered by plating a sample of salted squid, a value that cannot support the septicaemia observed in patients, but BacLight assay showed that more than 90% of the population was viable. In such a way, the hypothesis that an outbreak of foodborne bacteria in VBNC state could be underestimated on the basis of culturable cells has been demonstrated (Asakura et al., 2002).

In another outbreak in Japan, due to contamination of salted salmon roe by *E. coli* O157 in VBNC state, the authors demonstrated the resuscitation *in vivo* of VBNC bacteria and the maintaining of virulence in mice (Makino et al., 2000).

Such studies demonstrate that also treatments for food preservation have to be considered as a possible factor inducing pathogen bacteria into VBNC state. On the basis of such observations, the role of preservation treatments has to be investigated, in order to identify the critical points of a given procedure and make the adequate corrections.

9. Detection of VBNC bacteria

Actually, several systems for VBNC detection in water environments have been set up. Such systems are based on fluorescent staining, DNA hybridization and mRNA quantization.

Many fluorescent stains are used as indicators of metabolic activity (or viability), because they can accept the electrons flowing through cell respiratory chains. The most common are

known as **tetrazolium salts**, as 5-cyano-2,3-ditolyl tetrazolium chloride (CTC) and 2-(4-iodophenyl)-3-(4-nitrophenyl)-5-phenyl tetrazolium chloride (INT).

In recent years, a new differential staining assay, the BacLight® Live/Dead assay, has been developed. The assay allows to simultaneously count total and viable (metabolically active) cells, by using two nucleic acid stains, that is green-fluorescent SYTO® 9 stain and red-fluorescent propidium iodide stain. These stains differ in their ability to penetrate intact cell membranes. When used alone, SYTO® 9 stain labels both live and dead bacteria. In contrast, propidium iodide penetrates only bacteria with damaged membranes, reducing SYTO® 9 fluorescence when both dyes are present. Thus, live bacteria with intact membranes fluoresce green, while dead bacteria with damaged membranes fluoresce red.

CTC, INT and BacLight® Live/Dead assay are commonly used for the Direct Viable Count (DVC), by visualization of VBNC bacteria under fluorescence microscopy, one of the most widespread techniques for assessing bacterial viability (Rowan, 2011).

When DVC values are higher than culturable cells, obtained by CFU assay (that is when cell viability is higher than culturability), then it is assumed that bacteria have entered VBNC state.

Fluorescent antibodies have also been employed in DVC for VBNC detection of *E. coli* in recreational water (Zimmerman et al., 2009). In such technique, the sample is incubated in presence of yeast extract and nalidixic acid, which inhibits bacterial replication. Then, living cells increase their biomass, but cannot duplicate, so they elongate or enlarge. After addition of fluorescent antibody, fluorescing cells with altered morphology can be counted with epifluorescence microscopy. The difference between fluorescing cells, obtained by DVC, and culturable cells, obtained by CFU assay, indicates the presence of VBNC bacteria.

DNA hybridization is based on the identification of nucleic acid sequences that are specific for a given species, such as those present in ribosomes (16S and 23S rRNAs), by using a fluorescent known DNA sequence, called **probe**.

In **fluorescence in situ hybridization** (FISH), the hybridization between probe and nucleic acid sequence can be visualized in epifluorescence microscopy and used for DVC after incubation of the sample in a medium containing an antibiotic which prevents bacterial division. Comparison between DVC and culture count by miniaturized most-probable number (MPN) gives information about the presence of VBNC cells (Garcia-Armisen, 2004; Servais et al., 2009).

Flow cytometry has been successfully employed to gather information about cell viability, antigenic surface components, and the quantification of morphological variations of *V. parahaemolyticus* during entry into the VBNC state (Falcioni et al., 2008).

mRNA quantization is performed by quantitative reverse transcription polymerase chain reaction (qRT-PCR). Such technique is derived from the classic polymerase chain reaction (PCR), which allows the synthesis of huge quantities of a DNA sequence.

Dunaev et al. (2008) recently reported on the rapid and accurate quantification of VBNC pathogens in biosolids via monitoring and quantifying stress-related genes in *Salmonella* spp. using cDNA microarrays combined with qRT-PCR. Quantification of mRNA was correlated to cell viability and their ability to grow.

Other techniques, such as microradiography, have been proposed for VBNC identification, but they resulted to be time-consuming and quite expensive.

The above described detection methods, used in water environment, are not part of routinary food safety procedures, in that they are not simple enough, the equipment is too expensive and specialised technical personnel is necessary.

Moreover, it has to be considered that food could be quite difficult to investigate, because of the presence of heterogeneous compounds that could interfere with the molecular reaction required by a single detection technique.

However, new methods for VBNC detection should be designed, in terms of rapidity, sensitivity and ease of use.

10. Social and economical consequences of an outbreak

In any outbreak, beyond the health aspect, several social and economic implications are strictly associated to the emergence of a foodborne disease.

20% of illnesses are referable to known pathogens, but the remaining 80% is due to unspecified agents, intended as known agents not yet identified as cause of foodborne illness, agents known to be in food whose pathogenicity is not proven and unidentified agents.

As consequence, their traceability by traditional food safety methods is difficult.

In this regard, VBNC role in foodborne outbreaks has still to be defined.

In fact, the appearance of an outbreak has a critical effect on public opinion, generating fears that often become panic. The difficulty of a correct communication among health authorities, politics and people has been matter of study by World Health Organization, which published a manual containing the guidelines to overcome such problem.

Moreover, the difficulty to identify in a timely fashion the primary causes of the outbreak has a negative impact at economic level, because any foodstuff suspected to be infected is removed from the market and destroyed, with consequent economic losses.

The management of the recent outbreak of *E. coli* O104:H4 in Germany can be considered as a good source for several considerations.

At the onset of the outbreak, the identification of the aetiological agent was erroneous and only later it was demonstrated that the true cause was an enteroaggregative *E. coli strain,* able to synthesize *Shiga* toxins.

The source of infection was sprouted food produced by an organic farm in Bienenbüttel, Lower Saxony, Germany; local laboratories failed in finding the pathogen and correct identification was achieved later by a laboratory in North Rine-Westphalia.

Before the test results, German health authorities erroneously declared that the pathogen was present in cucumbers imported from Spain. Later, they admitted that the *E. coli* strain responsible of the outbreak was not found in cucumber samples they had analysed.

As consequence of the false information, Spanish exporters lost more than 200 millions of Euro per week. Political tension between Spain and Germany exploded on behalf of the European summit in Debrecen, Hungary, where the Spanish Minister of Agriculture publicly accused the German government of irresponsible behaviour. Spain, in fact, is the first world producer of cucumbers and more than 90% of its agricultural produce is exported in Europe. The destruction of vegetables and fruit had dramatic economic losses, because at that time Spain was already heavily hit by international financial crisis.

Moreover, Russia banned the import of fresh vegetable from Europe, with economic consequences which impacted on all the countries of European Union.

Another cause of diplomatic tension was a joint risk-assessment by European Food Safety Agency (EFSA)/European Centre for Disease Prevention and Control (ECDC), which identified a link between the German outbreak and a simultaneous haemolytic urea syndrome outbreak in France, caused by the same *E. coli* strain. The assessment indicated fenugreek seeds imported in Europe from Egypt as a possible source of contamination, even if in the same document it was stated that such hypothesis had to be truly demonstrated. Such behaviour induced the Egyptian Minister of Agriculture to comment such opinions as "sheer lies".

Several critical points about the German outbreak management emerge. First of all, standard procedures for the fast and unambiguous pathogen(s) identification are required, in order to alert health care points and allow them to plan the best response for people assistance.

Then, a real-time communication between health care points and national health authorities is needed, for continuous monitoring of the outbreak.

Health authorities have to communicate with politics, to plan the best way to inform people about the situation, giving appropriate information and warnings.

Further, communication among health authorities of the countries involved in the outbreak has to be promoted, to efficiently identify the causes of the outbreak and track it.

Communication, then, is the keystone for the best management of an outbreak. Appropriate procedures have to be established, with the creation of multilevel joint committees, formed by health and political authorities, which have to control quality and completeness of available information and manage subsequent communication to public opinion.

The importance of communication has been matter of the WHO Expert Consultation on Outbreak Communications held in Singapore on 2004. As result, a manual containing the guidelines to be followed has been published and is freely available (WHO, 2005).

11. International programs devoted to monitoring and tracking of foodborne pathogens

Surveillance of foodborne disease is a fundamental component of food safety systems and data are used for planning, implementing and evaluating public health policies. There is therefore a strong need to strengthen such systems, particularly for establishing whether VBNC foodborne pathogens may be responsible for an outbreak.

In this regard, the New Zealand Food Safety Authority has charged the Institute of Environmental Science and Research Limited to investigate the resuscitation of putative VBNC foodborne bacteria of significance to New Zealand. The final report concluded that some foodborne pathogen bacteria may become VBNC under certain conditions, that there may be no universal system for resuscitation of VBNC cells and that the phenomenon may be highly variable and bacterial species specific.

The difficulty in investigating the VBNC state and the related potential risk did not allow health authorities to establish guidelines for their detection.

Anyway, several efforts are carried out for surveillance of emerging foodborne diseases, creating new and interdisciplinary teams of research for data generation, collection and analysis.

In 2000, the Food and Agriculture Organization (FAO) of the United Nations and World Health Organization (WHO) started to expand their activities in the area of microbiological risk assessment to meet the increasing need for risk-based scientific advice and information and tools to undertake microbiological risk assessment. FAO and WHO coordinate their work in this area through the implementation of joint FAO/WHO meetings on microbiological risk assessment (JEMRA).

The activities of JEMRA can be categorised as follows:

- Generation of scientific information - risk assessments
- Elaboration of guideline documents
- Data collection and generation
- Use of risk assessment within a risk management framework
- Information and technology transfer

Moreover, WHO is promoting many programs and projects, by the creation of several worldwide collaborations involving technical existing structures, to give support, information and instructions on how to face an incoming outbreak.

The Global Alert and Response (GAR) is an integrated system for epidemics and other public health emergencies based on strong national public health systems and capacity and an effective international system for coordinated response.

The Global Outbreak Alert and Response Network (GOARN) is a technical collaboration of existing institutions and networks that pool human and technical resources for the rapid identification, confirmation and response to outbreaks of international importance. The Network provides an operational framework to link this expertise and skill to keep the international community constantly alert to the threat of outbreaks and ready to respond.

In 2011, United States President Obama signed into law the FDA Food Safety Modernization Act (FSMA). It aims to ensure the U.S. food supply is safe by shifting the focus of federal regulators from responding to contamination to preventing it. As a mandate, the FDA will launch a test of two different programs that they hope will help with locating the source of food contamination more quickly.

One program will track processed foods, and the other will trace raw fruits and vegetables. The program will focus on keeping more detailed records of food and the path it makes as it travels across the country.

A new project has been established at Centers for Disease Control and Prevention (CDC), in collaboration with the Food and Drug Administration (FDA), which is called the Food-borne Disease Active Surveillance Network or FoodNet.

The project consists of active surveillance for foodborne diseases and related epidemiologic studies designed to help public health officials better understand the epidemiology of foodborne diseases in the United States.

In the United States, using FoodNet data from 2000–2007, the Centers for Disease Control and Prevention estimated that 48 million people get sick, 127,839 were hospitalized and 3,037 people died.

Similarly, the European Food Safety Authority (EFSA) is the keystone of European Union (EU) risk assessment regarding food and feed safety and emerging risks, in close collaboration with national authorities.

Particularly, European food safety policy is to ensure a high level of protection of human health and consumers' interests in relation to food, taking into account diversity, including traditional products, whilst ensuring the effective functioning of the internal market.

Since 2000, the European Commission's guiding principle, primarily set out in its White Paper on Food Safety, is to apply an integrated approach from farm to table covering all sectors of the food chain, from feed production to transport and retail sale.

Moreover, the regional office for Europe of WHO supports countries in building capacity to manage food safety challenges in accordance with the WHO European Action Plan for Food and Nutrition Policy 2007–2012 and the WHO global strategy for food safety. The Action Plan is an important guide for policy-makers and health professionals that includes a wide range of actions in the area of food safety.

Despite the care by public health entities to track efficiently many known foodborne pathogens and identify the possibility of outbreaks, it is known that a consistent number of outbreaks is due to unknown pathogens. Such data may be due to microbes that are not proven to cause diseases as well as to VBNC state of known foodborne pathogen bacteria that, up to date, are undetectable.

12. Conclusions

Foodborne pathogens are the greatest threat to food safety. Despite the several efforts, much remains to be done to reach the national health objectives.

The first action to be taken is the enhancement of measures to reduce or prevent contamination in the food and to educate stakeholders more effectively about risks and prevention measures.

Whereas in developed countries food safety procedures and surveillance network monitor continuously the insurgence of possible emergencies, the situation is drastically different in developing countries, where many difficulties have to be considered.

In general, the lack of people education is the first problem, in that social habits as well as cultural factors may favour the spread of a foodborne infection.

Then, people education is the keystone needed to strongly limit the raise and diffusion of an outbreak within a community.

Moreover, a continuous support, in terms of updating of political authorities and retraining of health personnel, could contribute to an efficient action by the local governments.

Another step for developing countries is the creation of a permanent educational system that allows people, since infancy to adult age, to be informed about foodborne infections, their causes and ways of spreading. In this way, such information will become part of the community culture and will be transmitted to the future generations.

Overall, it is important to encourage the local formation of specialised personnel in healthcare sector for medical care, research, monitoring and communication.

Furthermore, it would be also desirable that international health authorities could take into account the focusing of VBNC state of foodborne bacteria. In fact, despite the studies performed up to date, several topics have to be clarified, that is *i*) the risk that food may induce VBNC state in contaminating foodborne pathogens; *ii*) wheter VBNC state can be considered as a resistance strategy against stress or a transitory condition which precedes cell death; *iii*) the factors triggering the VBNC state; *iv*) the factors triggering the resuscitation and *v*) the maintaining of the virulence.

In fact, in food field it is very important to fit into a global frame the effects of the various food parameters, as chemophysical characteristics, decontamination procedures and storage conditions and time. All these factors, in fact, may be source of multiple and subsequent stresses for contaminating pathogen bacteria and evaluating the possibility of onset of VBNC state is a primary need. In such sense, bioinformatics and predictive mathematical models could be a powerful tool to identify whether a specific association of different stresses acting on a bacterial population may induce the entry into VBNC state.

Unfortunately, up to date detection of VBNC foodborne pathogen bacteria in food is problematic.

New methods have been proposed for VBNC detection, but they are not satisfying in that difficult to use as routinary procedures.

Therefore, innovative detection procedures, in terms of ease of use and rapidity of analysis, are urgently needed to allow an efficient monitoring and tracking of foodborne VBNC to prevent outbreaks in today's and future global market.

13. References

Asakura, H., Makino, S.I., Takagi, T., Kuri, A., Kurazano, T., Watarai M. & Shirahata, T. (2002). Passage in Mice Causes a Change in the Ability *of Salmonella enterica* Serovar Oranienburg to Survive NaCl Osmotic Stress: Resuscitation from the Viable but Non-culturable State. *FEMS Microbiology Letters*, Vol.212, No.1, (June 2002), pp. 87-93, ISSN 0378-1097

Asakura, H., Kawamoto, K., Haishima, Y., Igimi, S., Yamamoto, S. & Makino, S.I. (2008). Differential Expression of the Outer Membrane Protein W (OmpW) Stress Response in Enterohaemorrhagic *Escherichia coli* O157:H7 Corresponds to the Viable but Non-culturable State. *Research in Microbiology*, Vol.159, No.9-10, (November-December 2008), pp. 709-717, ISSN 0923-2508

Baffone, W., Citterio, B., Vittoria, E., Casaroli, A., Campana, R., Falzano, L. & Donelli, G. (2003). Retention of Virulence in Viable but Nonculturable Halophilic *Vibrio* spp. *International Journal of Food Microbiology*, Vol.89, No.1, (December 2003), pp. 31-39, ISSN 0168-1605

Berger, F., Morellet, N., Menu, F. and Potier, P. (1996). Cold Shock and Cold Acclimation Proteins in the Psychrotrophic Bacterium *Arthrobacter globiformis* SI55. *Journal of Bacteriology*, Vol.178, No.11, (June 1996), pp. 2999-3007, ISSN 0021-9193

Besnard, V., Federighi, M., Declerq, E., Jugiau, F. & Cappelier, J.M. (2002). Environmental and Physico-chemical Factors Induce VBNC State in *Listeria monocytogenes*. *Veterinary Research*, Vol.33, No.4, (July-August 2002), pp. 359-370, ISSN 0928-4249

Beumer, R.R., de Vries, J. & Rombouts, F.M. (1992). *Campylobacter jejuni* Non-culturable Coccoid Cells. *International Journal of Food Microbiology*, Vol.15, No.1-2, (January-February 1992), pp. 153-163, ISSN 0168-1605

Bore, E., Langsrud, S., Langsrud, Ø., Rode, T.M. & Holck, A. (2007). Acid-shock Responses in *Staphylococcus aureus* Investigated by Global Gene Expression Analysis. *Microbiology*, Vol.153, No.7, (July 2007), pp. 2289-2303, ISSN 1350-0872

Cappelier, J.M., Magras, C., Jouve, J.L. & Federighi, M. (1999a). Recovery of Viable but Nonculturable *Campylobacter jejuni* Cells in Two Animal Models. *Food Microbiology,* Vol.16, No.4, (August 1999), pp. 375-383, ISSN 0740-0020

Cappelier, J.M., Minet, J., Magras, C., Colwell, R.R. & Federichi, M. (1999b). Recovery in Embryonated Eggs of Viable but Nonculturable *Campylobacter jejuni* Cells and Maintenance of Ability to Adhere to HeLa Cells after Resuscitation. *Applied and Environmental Microbiology*, Vol.65, No.11, (November 1999), pp. 5154-5157, ISSN 0099-2240

Cappelier, J.M., Besnard, V., Roche, S., Garrec, N., Zundel, E., Velge, P. & Federighi, M. (2005). Avirulence of Viable but Non-culturable *Listeria monocytogenes* Cells Demonstrated by in Vitro and in Vivo Models. *Veterinary Research*, Vol.36, No.4, pp. 589-599, (July-August 2005), ISSN 0928-4249

Cappelier, J.M., Besnard, V., Roche, S.M., Velge, P. & Federighi, M. (2007). Avirulent Viable but Non-culturable Cells of *Listeria monocytogenes* Need the Presence of an Embryo to be Recovered in Egg Yolk and Regain Virulence after Recovery. *Veterinary Research*, Vol.38, No.4, (July-August 2007), pp. 573-583, ISSN 0928-4249

Colwell, R.R., Brayton, P.R., Herrington, D., Tall, B.D., Huq, A. & Levine M.M. (1996). Viable but Nonculturable *Vibrio cholerae* O1 Revert to a Culturable State in Human Intestine. *World Journal of Microbiology and Biotechnology*, Vol.12, No.1, (January 1996), pp. 28-31, ISSN 0959-3993

Cook, K.L. & Bolster, C.H. (2007). Survival of *Campylobacter jejuni* and *Escherichia coli* in Groundwater During Prolonged Starvation at Low Temperatures. *Journal of Applied Microbiology*, Vol.103, No.3, (September 2007), pp. 573-583, ISSN 1364-5072

Csonka, L.N. (1989). Physiological and Genetic Responses of Bacteria to Osmotic Stress. *Microbiological Reviews*, Vol.53, No.1, (March 1989), pp. 121-147, ISSN 0146-0749

Day, A.P. & Oliver, J.D. (2004). Changes in Membrane Fatty Acid Composition during Entry of *Vibrio vulnificus* into the Viable but Nonculturable State. *The Journal of Microbiology*, Vol.42, No.2, (June 2004), pp. 69-73, ISSN 1225-8873

Dunaev, T., Alanya, S. & Duran, M. (2008). Use of RNA-based Genotypic Approaches for Quantification of Viable but Nonculturable *Salmonella* sp. in Biosolids. *Water Science and Technology*, Vol.58, No.9, (n.d.), pp. 1823-1828, ISSN 0273-1223

Eberl, L., Givskov, M., Sternberg, C., Møller, S., Christiansen, G. & Molin, S. (1996). Physiological Responses of *Pseudomonas putida* KT2442 to Phosphate Starvation. *Microbiology*, Vol.142, No.1, (January 1996), pp. 155-163, ISSN 1350-0872.

Erickson, M.C., Webb, C.C., Diaz-Perez, J.C., Phatak, S.C., Silvoy, J.J., Davey, L., Payton, A.S., Liao, J., Ma, L. & Doyle M.P. (2010). Surface and Internalized *Escherichia coli* O157:H7 on Field-Grown Spinach and Lettuce Treated with Spray-Contaminated Irrigation Water. *Journal of Food Protection*, Vol.73, No.6, (June 2010), pp. 1023-1029, ISSN 0362-028X

Falcioni, T., Papa, S., Campana, R., Manti, A., Battistelli, M. & Baffone, W. (2008). State Transitions of *Vibrio parahaemolyticus* VBNC Cells Evaluated by Flow Cytometry. *Cytometry Part B (Clinical Cytometry)*, Vol.74, No.5, (September 2008), pp. 272-281, ISSN 1552-4949

Federighi, M., Tholozan, J.L., Cappelier, J.M., Tissier, J.P. & Jouve, J.L. (1998). Evidence of Non-coccoid Viable but Non-culturable *Campylobacter jejuni* Cells in Microcosm Water by Direct Viable Count, CTC-DAPI Double Staining, and Scanning Electron Microscopy. *Food Microbiology*, Vol.15, No.5, (October 1998), pp. 539-550, ISSN 0740-0020

Freestone, P.P.E., Haigh, R.D., Williams, P.H. & Lyte, M. (1999). Stimulation of Bacterial Growth by Heat-stable, Norepinephrine-induced Autoinducers. *FEMS Microbiology Letters*, Vol.172, No.1, (March 1999), pp. 53-60, ISSN 0378-1097

Gage, J. & Neidhardt, F.C. (1993). Modulation of the Heat Shock Response by One-carbon Metabolism in *Escherichia coli*. *Journal of Bacteriology*, Vol.175, No.7, (April 1993), pp. 1961-1970 ISSN 0021-9193

Garcia-Armisen, T. & Servais, P. (2004). Combining Direct Viable Counts (DVC) and Fluorescent in Situ Hybridization (FISH) to Enumerate Viable *Escherichia coli* in Rivers and Wastewaters. *Water Science and Technology*, Vol.50, No.1, (n.d.), pp. 271-275, ISSN 0273-1223

Ghezzi, J.I. & Steck, T.R. (1999). Induction of the Viable but Nonculturable Conditions in *Xanthomonas campestris* pv. *campestris* in Liquid Microcosms and Sterile Soil. *FEMS Microbiology Ecology*, Vol.30, No.3, (November 1999), pp. 203-208, ISSN 0168-6496

Givskov, M., Eberl, L., Møller, S., Poulsen, L.K., Molin, S. (1994). Responses to Nutrient Starvation in *Pseudomonas putida* KT2442: Analisys of General Cross-protection, Cell Shape, and Macromolecular Content. *Journal of Bacteriology*,. 1994 Vol.176, No.1, (January 1994), pp. 7-14, ISSN 0021-9193

Gourmelon, M., Cillard, J. & Pommepuy, M. (1994) Visible Light Damage to *Escherichia coli* in Seawater: Oxidative Stress Hypothesis. *Journal of Applied Microbiology*, Vol.77, No.1, (July 1994), pp. 105-112, ISSN 1364-5072

Gunasekera, T.S., Sørensen, A., Attfield, P.V., Sørensen, S.J. & Veal, D.A. (2002). Inducible Gene Expression by Nonculturable Bacteria in Milk after Pasteurization. *Applied*

and Environmental Microbiology, Vol.68, No.4, (April 2002), pp. 1988-1993, ISSN 0099-2240

Hett, E.C., Chao, M.C., Steyn, A.J., Fortune, S.M., Deng, L.L. & Rubin, E.J. (2007). A Partner for the Resuscitation-promoting Factors of *Mycobacterium tuberculosis*. *Molecular Microbiology*, Vol.66, No.3, (November 2007), pp. 658-668, ISSN 0950-382X

Jones, P.G. & Inouye, M. (1996). RbfA, a 30 S Ribosomal Binding Factor, is a Cold-Shock Protein Whose Absence Triggers the Cold-Shock Response. *Molecular Microbiology*, Vol.21, No.6, (September 1996), pp. 1207-1218, ISSN 0950-382X

Kana, B.D., Gordhan, B.G., Downing, K.J., Sung, N., Vostroktunova, G., Machowski, E.E., Tsenova, L., Young, M., Kaprelyants, A., Kaplan, G. & Mizrahi, V.l. (2008). The Resuscitation-promoting Factors of *Mycobacterium tuberculosis* are Required for Virulence and Resuscitation from Dormancy but are Collectively Dispensable for Growth in Vitro. *Molecular Microbiology*, Vol.67, No.3, (February 2008), pp. 672-684, ISSN 0950-382X

Kawasaki, S., Watamura, Y., Ono, M., Watanabe, T., Takeda, K. & Niimura, Y. (2005). Adaptive Responses to Oxygen Stress in Obligatory Anaerobes *Clostridium acetobutylicum* and *Clostridium aminovalericum*. *Applied and Environmental Microbiology*, Vol.71, No.12, (December 2005), pp. 8442-8450, ISSN 0099-2240

Kelly, S.M., Pitcher, M.C.L., Farmery, S.M. & Gibson, G.R. (1994). Isolation of *Helicobacter pylori* from Feces of Patients with Dyspepsia in the United Kingdom. *Gastroenterology*, Vol.107, No.6, (December 1994), pp. 1671-1674, ISSN 0016-5085

Kurath, G. & Morita, R.Y. (1983). Starvation-Survival Physiological Studies of a Marine *Pseudomonas* sp. *Applied and Environmental Microbiology*, Vol.45, No.4, (April 1983), pp. 1206-1211, ISSN 0099-2240

Lleò, M.M., Pierobon, S., Tafi, M.C., Signoretto, C. & Canepari, P. (2000). mRNA Detection by Reverse Transcription-PCR for Monitoring Viability over Time in an *Enterococcus faecalis* Viable but Nonculturable Population Maintained in a Laboratory Microcosm. *Applied and Environmental Microbiology*, Vol.66, No.10, (October 2000), pp. 4564-4567, ISSN 0099-2240

Lleò, M.M.; Bonato, B.; Tafi, M.C.; Signoretto, C.; Boaretti, M. & Canepari, P. (2001). Resuscitation Rate in Different Enterococcal Species in the Viable but Non-culturable State. *Journal of Applied Microbiology*, Vol.91, No.6, (December 2001), pp. 1095-1102, ISSN 1364-5072,

Loftis, A.D.; Priestley, R.A. & Massung, R.F. (2010). Detection of *Coxiella burnetii* in Commercially Available Raw Milk from the United States. *Foodborne Pathogens and Disease*, Vol.7, No.12, (December 2010), pp. 1453-1456, ISSN 1535-3141

Makino, S.I.; Kii, T.; Asakura, H.; Shirahata, T.; Ikeda, T.; Takeshi, K. & Itoh, K. (2000). Does Enterohaemorrhagic *Escherichia coli* O157:H7 Enter the Viable but Nonculturable State in Salted Salmon Roe? *Applied and Environmental Microbiology*, Vol.66, No.12, (December 2001), pp. 5536-5539, ISSN 0099-2240

Matin, A. (1991). The Molecular Basis of Carbon-Starvation-Induced General Resistance in *Escherichia coli*. *Molecular Microbiology*, Vol.5, No.1, (January 1991), pp. 3-10, ISSN 0950-382X

McDougald, D. & Kjelleberg, S. (2006). Adaptive Responses of Vibrios. In: *The biology of vibrios*, F. L Thompson,. B. Austin & J. G. Swings (Eds.), 133-155, ASM Press, ISBN 1555813658, Herndon, Virginia, United States

McGovern, V.P. & Oliver, J.D. (1995). Induction of Cold Responsive Proteins in *Vibrio vulnificus*. *Journal of Bacteriology*, Vol.177, No.14, (July 1995), pp. 4131-4133, ISSN 0021-9193

Millet, V., & Lonvaud-Funel, A. (2000). The viable but non-culturable state of wine microorganisms during storage. *Letters in Applied Microbiology*, Vol.30, No.2, (February 2000), pp.136-141, ISSN 0266-8254

Morita, R.Y. (1982). Starvation-Survival of Heterotophs in the Marine Environment. *Advances in Microbial Ecology*, Vol.6, (n.d.), pp. 117-198, ISSN 0147-4863

Morton, D. & Oliver, J.D. (1994). Induction of Carbon Starvation Proteins in *Vibrio vulnificus*. *Applied and Environmental Microbiology*, Vol.60, No.10, (October 1994), pp. 3653-3659, ISSN 0099-2240

Mukamolova, G.V., Kaprelyants, A.S., Young D.I., Young, M. & Kell, D.B. (1998a). A bacterial cytokine. *Proceedings of National Academy of Sciences*, Vol.95, No.15, (July 1998), pp. 8916-8921, ISSN 0027-8424

Mukamolova, G.V.; Yanopolskaya, N.D.; Kell, D.B. & Kaprelyants, A.S. (1998b). On resuscitation from the dormant state of *Micrococcus luteus*. Antonie Van Leeuwenhoek, Vol.73, No.3, (Apr 1998), pp. 237-243, ISSN 0003-6012

Nakashima, K., Kanamaru, K., Mizuno, T. & Horikoshi, K. (1996). A Novel Member of the *cspA* Family of Genes that is Induced by Cold Shock in *Escherichia coli*. *Journal of Bacteriology*, Vol.178, No.10, (May 1996), pp. 2994-2998, ISSN 0021-9193

Nesbakken, T. (2007). Pig Herds Free from Human Pathogenic *Yersinia enterocolitica*. *Emerging Infectious Diseases*, Vol.13, No.12, (December 2007), pp. 1860-1864, ISSN 1080-6059

Nicolò, M.S., Gioffre`, A., Carnazza, S., Platania, G., Di Silvestro, I. and Guglielmino, S.P.P. (2011) Viable but Nonculturable State of Foodborne Pathogens in Grapefruit Juice: a Study of Laboratory. *Foodborne Pathogens and Disease*, Vol.8, No.1, (January 2011), pp. 11-17, ISSN 1535-3141

Nyström, T., Olsson, R.M. & Kjelleberg, S. (1992). Survival, Stress Resistance, and Alterations in Protein Expression in the Marine *Vibrio* sp. Strain S14 during Starvation for Different Individual Nutrients. *Applied and Environmental Microbiology*, Vol.58, No.1, (January 1992), pp. 55-65, ISSN 0099-2240

Oliver, J.D. (2000). Problems in Detecting Dormant (VBNC) Cells and the Role of DNA Elements in This Response, In: *Tracking Genetically-engineered Microorganisms*, J.K. Jansson, J.D. van Elsas, and M.J. Bailey (eds.), 1-15, Landes Biosciences, ISBN 978-1-58706-009-0, Georgetown, Texas, USA.

Oliver, J.D. (2005). Viable but Nonculturable Bacteria in Food Environments. In: *Food Borne Pathogens: Microbiology and Molecular Biology*, P.M. Fratamico & A.K. Bhunia (Eds.), 99-112, Horizon Scientific Press, ISBN: 978-1-898486-52-7, Norfolk, United Kingdom

Oliver, J.D. & Bockian, R. (1995). In Vivo Resuscitation, and Virulence Towards Mice, of Viable but Nonculturable Cells of *Vibrio vulnificus*. *Applied and Environmental Microbiology*, Vol.61, No.7, (July 1995), pp. 2620-2623, ISSN 0099-2240

Porter, J., Edwards, C. & Pickup, R.W. (1995). Rapid Assessment of Physiological Sstatus in *Escherichia coli* Using Fluorescent Probes. *Journal of Applied Bacteriology*, Vol.79, No.4, (October 1995), pp. 399-408, ISSN 1364-5072

Quirós, C., Herrero, M., Garcia, L.A. & Diaz, M. (2009). Quantitative Approach to Determining the Contribution of Viable-butnonculturable Subpopulations of

Malolactic Fermentation Processes. *Applied and Environmental Microbiology* Vol.75, No.9, (May 2009), pp. 2977-2981, ISSN 0099-2240

Reissbrodt, R., Rienaecker, I., Romanova, J.M., Freestone, P.P.E., Haigh, R.D., Lyte, M., Tschäpe, H. & Williams, P.H. (2002). Resuscitation of *Salmonella enterica* Serovar Typhimurium and Enterohemorrhagic *Escherichia coli* from the Viable but Nonculturable State by Heat-stable Enterobacterial Autoinducer. *Applied and Environmental Microbiology*, Vol.68, No.10, (October 2002), pp. 4788-4794. ISSN 0099-2240

Rockabrand, T.A., Arthur, T., Korinek, G., Livers, K. & Blum, P. (1995). An Essential Role for the *Escherichia coli* DnaK Protein in Starvation-induced Thermotolerance, H_2O_2 Resistance, and Reductive Division. *Journal of Bacteriology*, Vol.177, No.13, (July 1995), pp. 3695-3703, ISSN 0021-9193

Rowan, N.J. (2011). Defining Established and Emerging Microbial Risks in the Aquatic Environment: Current Knowledge, Implications, and Outlooks. *International Journal of Microbiology*, Vol.2011, (n.d.), pp. 1-15, ISSN 1687-918X

Servais, P., Prats, J., Passerat, J. & Garcia-Armisen, T. (2009). Abundance of Culturable versus Viable *Escherichia coli* in Freshwater. *Canadian Journal of Microbiology*, Vol.55, No.7, (July 2009), pp. 905-909, ISSN 0008-4166

Shleeva, M., Mukamolova, G.V., Young, M., Williams, H.D. & Kaprelyants, A.S. (2004). Formation of 'Non-culturable' Cells of *Mycobacterium smegmatis* in Stationary Phase in Response to Growth under Suboptimal Conditions and Their Rpf-mediated Resuscitation. *Microbiology*, Vol.150, No.6, (June 2004), pp. 1687-1697, ISSN 1350-0872

Signoretto, C., Lleò, M.M., Tafi, M.C. & Canepari, P. (2000). Cell Wall Chemical Composition of *Enterococcus faecalis* in the Viable but Nonculturable State. *Applied and Environmental Microbiology*, Vol.66, No.5, (May 2000), pp. 1953-1959, ISSN 0099-2240

Signoretto, C., Lleò, M.M. & Canepari, P. (2002). Modification of the Peptidoglycan of *Escherichia coli* in the Viable but Nonculturable State. *Current Microbiology*, Vol.44, No.2, (February 2002), pp. 125-131, ISSN 0343-8651

Spence, J., Cegielska, A. & Georgopoulos, C. (1990). Role of *Escherichia coli* Heat Shock Proteins DnaK and HtpG (C62.5) in Response to Nutritional Deprivation. *Journal of Bacteriology*, Vol.172, No.12, (December 1990), pp. 7157-7166, ISSN 0021-9193

Sperandio, V., Torres, A.G., Jarvis, B., Nataro, J.P. & Kaper, J.B. (2003). Bacteria–host Communication: the Language of Hormones. *Proceedings of the National Academy of Sciences*, Vol.100, No.15, (July 2003), pp. 8951–8956, ISSN 0027-8424

Sun, F., Chen, J., Zhong, L., Zhang, X.-H., Wang, R., Guo, Q. & Dong, Y. (2008). Characterization and Virulence Retention of Viable but Nonculturable *Vibrio harveyi*. FEMS Microbiology Ecology , Vol.64, No.1, (April 2008), pp. 37–44, ISSN 0168-6496

Terpeluk, C., Goldmann, A., Bartmann, P. & Pohlandt, F. (1992). *Plesiomonas shigelloides* Sepsis and Meningoencephalitis in a Neonate. *European Journal of Pediatrics*. Vol.151, No.7, (July 1992), pp. 499-501, ISSN 0340-6199

Thomas, J.E., Gibson, G., Darboe, M., Dale, A. & Weaver L.T. (1992). Isolation of *Helicobacter pylori* from Human Faeces. *Lancet*, Vol.340, No.8829, (November 1992), pp. 1094-1095, ISSN 0140-6736

WHO. (2005). WHO Outbreak Communication Guidelines. Available at: www.who.int/infectious-disease-news/IDdocs/whocds200528/whocds200528en.pdf.

Xu, H.S., Roberts, N., Singleton, F.L., Attwell, R.W., Grimes, D.J. & Colwell, R.R. (1982). Survival and Viability of Nonculturable *Escherichia coli* and *Vibrio cholerae* in the Estuarine and Marine Environment. *Microbial Ecology*, Vol.8, No.4, (December 1982), pp. 313-323, ISSN: 0095-3628

Yaron, S. & Matthews, K. (2002). A Reverse Transcriptase-polymerase Chain Reaction Assay for Detection of Viable *Escherichia coli* O157:H7: Investigation of Specific Target Genes. *Journal of Applied Microbiology*, Vol.92, No.4, (April 2002), pp. 633-640, ISSN 1364-5072

Zimmerman, A.M., Rebarchik, D.M., Flowers, A.R., Williams, J.L. & Grimes, D.J. (2009). *Escherichia coli* Detection Using mTEC Agar and Fluorescent Antibody Direct Viable Counting on Coastal Recreational Water Samples. *Letters in Applied Microbiology*, Vol.49, No.4, (October 2009), pp. 478–483, ISSN 0266-8254

10

Waste Minimization for the Safe Use of Nanosilver in Consumer Products – Its Impact on the Eco-Product Design for Public Health

K. W. Lem[1,7] et al.[*]
[1]*Department of Physics, MTSE Program, New Jersey Institute of Technology, NJ,*
[7]*Nanobiz, LLC, NJ, USA*

1. Introduction

Winslow (1920) has defined the meaning of public health as "the science and art of preventing disease, prolonging life and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals". Thus, the focus of public health intervention is the improvement of health and quality of life through the prevention and treatment of disease, and promotion of healthy behaviors. Promotion of hand washing is a typical common public health practice to prevent the spread of "unwanted" diseases (Samuel et al., 2005).

With the rapid increase on applications for nanoparticle silver, its potential impact on public health has become a critical issue. Nanosized silver can be made with different shapes such as particles, wires, and rods. Silver nanoparticles (AgNPs; many other names such as nanosilver (nAg) and colloidal silver) have already been used in everyday consumer

[*] S-H. Hsu[2], D. S. Lee[3], Z. Iqbal[4], S. Sund[5], S. Curran[6], C. Brumlik[7], A. Choudhury[7], D. S-G. Hu[8], N. Chiu[9], R. C. Lem[10] and J. R. Haw[11,#]
1 Department of Physics, MTSE Program, New Jersey Institute of Technology, NJ, USA,
2 Institute of Polymer Science and Engineering, National Taiwan University, Taipei, Taiwan,
3 Department of Chemical Engineering, Chonbuk National University, Chonju, Korea,
4 Department of Chemistry and Environ Science, New Jersey Institute of Technology, NJ, USA,
5 Nygard Consulting, LLC, NJ, USA,
6 Boston Scientific, MA, USA,
7 Nanobiz, LLC, NJ, USA,
8 Dept of Mat Sci and Eng, National Taiwan University Science and Technology, Taipei, Taiwan,
9 UMDNJ-New Jersey Medical School, Newark, NJ, USA,
10 Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA,
11 Department of Materials Chemistry and Engineering, Konkuk University, Seoul, Korea.
[#] Corresponding Author

products requiring broad spectrum antibiotic performance because of their enormous surface area and reactivity. Faunce & Watal (2010) recently analyzed international regulatory issues for medical and domestic use in the United States, European Union, United Kingdom, and Australia. They found that despite the numerous studies reported in recent decades, many scientists are still uncertain of its safety. Very recently, Powers (2010) showed positive that Ag+ and AgNPs are developmental neurotoxicants in vitro and in vivo. Therefore, there is a need to conduct a study to identify a global landscape of AgNPs, their products, and their manufacturers. A market-based intellectual property (IP) study has been conducted to examine the current global patent landscape of companies using AgNPs in their consumer product development and production from 1980 to 2010 (Lem et al., 2012). Information on materials, compositions, formulation, manufacturing processes, and ultimate application were extracted using a "two-stage" stage-gate process using the IP activity in the use of nanosilver in consumer products. The two stages studied were commercial and consumer products. In the first stage for AgNPs and AgNPs-based commercial products, Lem et al. (2012) reported that there were 7,422 patent families from January 1, 1980 to December 31, 2010. In the second stage for AgNPs-based consumer products, 932 patent families from January 1, 1980 for to December 31, 2010 were found.

Korea, China and the USA were found to be the major players in AgNPs and AgNPs-based commercial and consumer products. Korea has been the leader and consumer products containing nanosilver are even sold on the streets. Due to its enormous surface reactivity, nanosilver has already found utility in everyday products that require antibiotic performance, such as materials that contact food, textiles and fabrics, appliances, consumer products, children's toys, infant products, 'health' supplements, cosmetics and pharmaceuticals (Chaloupka et al., 2010). Thus, safety has become a potentially critical issue for companies that make products containing nanosilver. For product stewardship to their customers, suppliers must address environmental health and safety issues in terms of real risks, perceptual risks, and government regulation.

In this chapter, we evaluate consumer products containing nanosilver especially in Korea using an IP landscaping study. We examine the recent literature regarding the effect of silver and nanosilver on public health from a clinical medicine perspective. We then evaluate the concept of waste generation and waste minimization in the material life cycle. We examine the flow of silver and nanosilver in the life cycle of food metabolism to identify ways to minimize potential adverse effects of nanosilver and to provide concepts of eco-products to minimize exposure to nanosilver for public health. Coupling the recent literature and the IP landscaping study, together with Design for Lean Six Sigma – Green (DFLSS-G) and TRIZ, we demonstrate the concept of waste minimization by controlling the release of nanosilver (Lem et al., 2011; Liu JG et al., 2010) in eco-product design.

2. Manufacturing commercial AgNPs products

Nanoproducts can be produced in two ways: top-down or bottom-up. A top-down approach is essentially tearing down of a device to gain insight into its components, materials and compositions (e.g., ball milling). A bottom-up approach is the piecing together of materials to give rise to components and finally to build a device (e.g., chemical precipitation).

Figure 1 gives a roadmap of top-down and bottom-up development of nanoproducts under a value chain of "material – properties – processing – structure – performance – applications". This roadmap helps to identify the unmet needs from materials to/from applications in nanoproduct development and manufacture.

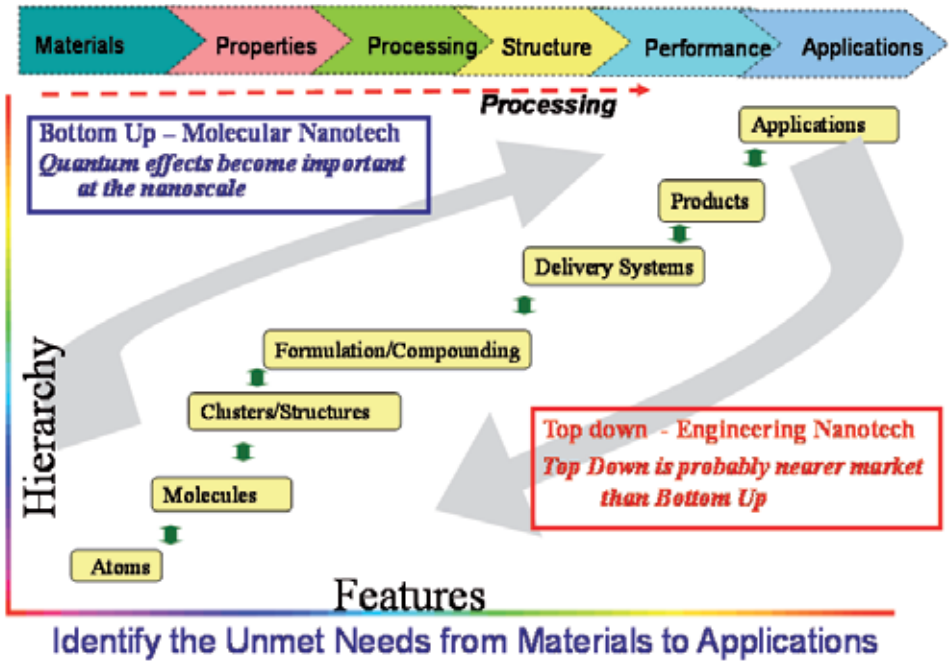


Fig. 1. Top-Down and Bottom-Up Development [Lem et al., 2009; Brauer et al., 2009)]

Commercial AgNPs products are typically produced via a bottom-up process, while analysis of environmental impact by AgNPs products is a top-down process (Tolaymat et al., 2010).

In order to use nanosilver for different applications, various chemical and physical methods are used to synthesize the nanosilver. One of the best recent overviews is the U.S. EPA's report by EI-Badawy et al. (2010). Chemical methods usually require a silver salt precursor, a reducing agent (formaldehyde, glucose or hydrogen peroxide.), a solvent, a stabilizer and a capping agent. Silver nitrate is a commonly used precursor to produce silver nanoparticles (Mousa & Linhart, 2010). Alternatively, the wet/dry electroplating method (Yang et al, 2006), electric spark bombardment (Koecher et al., 2009), and other mechanical techniques (polishing and grinding) are used as physical methods to produce nanoparticles.

Fabrication of nanosilver in the nanofiber form is usually done by electro spinning techniques. Other methods are being explored in applications to wound dressing and textile or personal care products.

Coatings have been used to employ the antibacterial property of nanosilver in biomedical devices or electronic devices. The nanosilver coating techniques include technologies such as spraying , chemical vapor deposition, dipping, printing knife-coating, transfer coating, and spin coating (Koecher et al., 2009).

Furthermore it is known that functionalizing the nanosilver with different functional groups/moieties can make nanosilver suitable for specific applications like biomedical devices, dye-doped particles, biomedical imaging, *etc.* One interesting example uses glycosaminoglycan conjugated nanosilver particles for biomedical devices (Mousa & Linhart, 2010). Anti-coagulation is another active area for research where a core of metal silver is coated with a surface layer of silver oxide (Zhu et al., 2002; Zhu & Zhu, 2004). One method of increasing sterilizing power and antibacterial property of nanosilver is by mixing it with nanosulfur compounds (Oh, 2002). Coated nanoparticles are produced by precipitating nanoparticle cores from two aqueous reactant solutions in a microemulsion, and adding a coating agent to coat the cores (Tan, 2003).

3. The landscaping of consumer products containing nanosilver

To have a better global picture of nanosilver in commercial and consumer products, Lem et al. (2012) have conducted an exhaustive intellectual property (IP) search study using a "two stages" stage-gate process. This search is divided into three parts based on timelines.

Part A. Before 1980 "pre nano": In this the search is done to identify patent documents which have a mention of nano sized silver material before nanotechnology became an establish field of science and technology.

Part B: Jan 1980 to Jul 2008: Time prior to EPA expressing their concern about toxicological uncertainty as a consequence of using nano scale silver.

Part C: Aug 2008 to Dec 2010: This is the time after the EPA has started showing concern towards toxicological uncertainties as a consequence of using nano scale silver. The search is terminated 2010 to avoid complications of handling data for the incomplete year 2011.

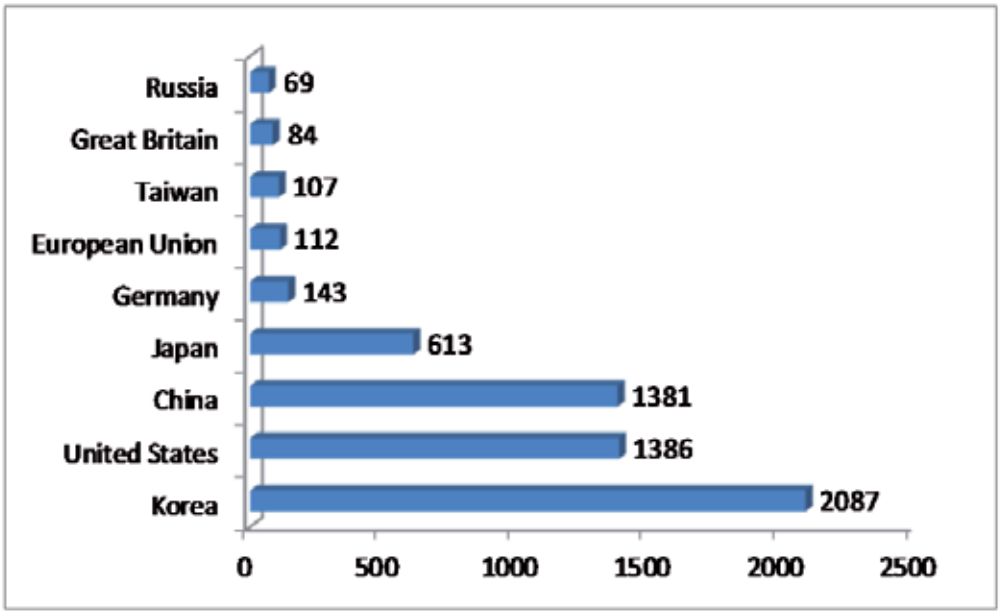


Fig. 2. AgNPs Patent Filing Countries (Adopted from Lem et al., 2012)

The details of the search methodology and results are not the subjects of this article. They have been reported recently by Lem et al. (2012). Figure 2 shows the countries researching AgNPs. Figure 2 shows that out of all the major patenting authorities, South Korea is leading with 2087 publications. The U.S. is second and China is third. These three countries have about 80% of the total patent publications. It is important to note that Chinese patent numbers are over-weighted by their Chinese Utility Model Patents that only protect the shape and/or structure of a product for 10 years and are not subject to substantive examination. The International Bureau or WIPO's Patent Cooperation Treaty (PCT) is a parallel patenting route that covers most of the world's patenting countries.

Due to its antiseptic and biocidal nature, more than 50% patenting activity of AgNPs is seen in the area of pharmaceuticals, healthcare, consumer goods, preservatives, sterilization, and water treatment. Using the patent landscape analysis, we identified relevant patents for pharmaceutical, healthcare, and consumer goods patents. Intensive IP analysis, such as broadness of claims, details of office actions can help in identifying the share of the market that can be attributed to current and future technology. Active companies in the technology space were identified as they can influence the market. For example, LG Household and Health Care Ltd (health care) and L'Oreal (cosmetics) are amongst the major assignees in the space of pharma and consumer goods patents.

Several observations were found in the 932 patent families from January 1, 1980 to December 31st, 2010 that was related to AgNP in consumer products. These 932 patent publications were reviewed to distil information with respect to:

1. Nanosilver size or percentage loading,
2. Nanosilver form or geometry,
3. Material composition, and
4. Application area and products.

Among these 932 publications, cosmetics, personal care, medical, and health care occupy more than 70% of the application areas as shown in Figure 3.

In order to analyze the data in more detail, we defined four different classes of formulation to examine the role of the concentration of nanosilver (e.g., dose) used in the formulations on these applications:

1. Trace dose, less than 0.01 wt. %.
2. Low dose, 0.01 to 1.0 wt. %.
3. Medium dose, 1.0 to 10 wt. %.
4. High dose, greater than 10 wt. %.

Clearly trace dose formulation found applications in foods, drinking water, drugs, facial mask, cream, wound care, gel, and textiles (Lem et al., 2012). At low and medium doses, the formulations found applications in household products materials, medical, personal care and cosmetics. Typical medical applications include catheter, endotracheal tube, and subcutaneous central venous port; pacemakers, prosthetic heart valves, prosthetic joints, voice prostheses, contact lenses, stents, heart valves, penile implants, small or temporary joint replacement, urinary dilators, cannulae, and intrauterine devices; catheter lock, needle, luer-lok (medical device) connectors, needleless connectors, clamps, forceps, scissors, skin hooks, tubing, needles, retractors, sealer, drills, chisel, rasps, surgical instruments, dental instruments, intravenous tubes, breathing tubes, dental water line, dental drain tubes, feeding tubes, bandages, wound dressings, orthopedic implants, and saws (Lem et al., 2012).

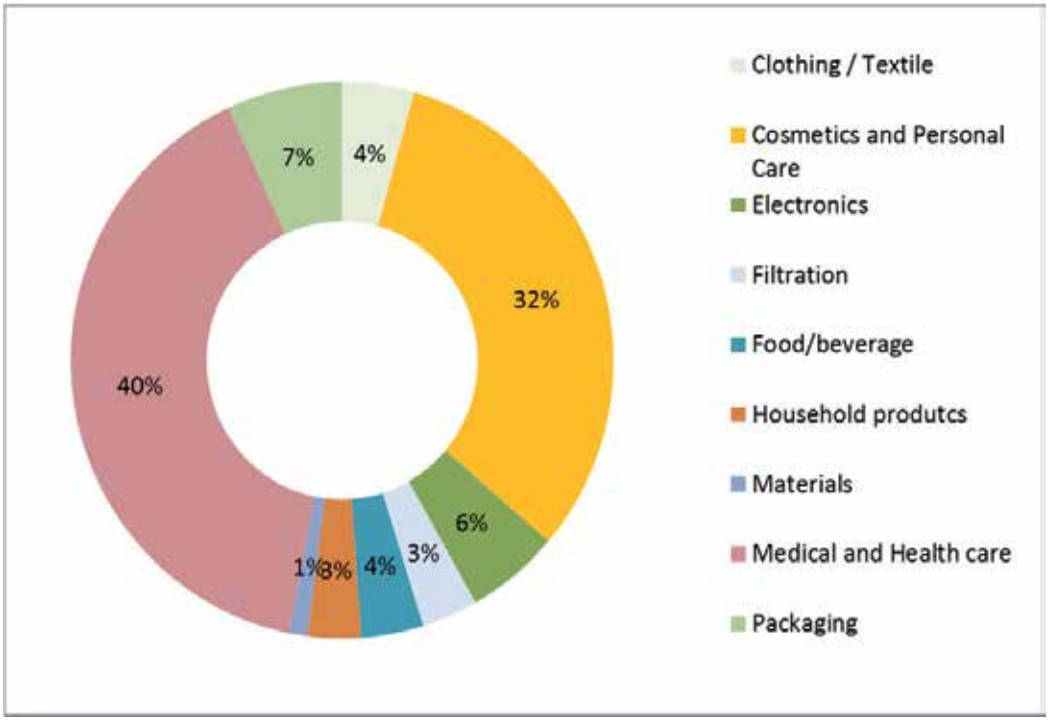


Fig. 3. Nanosilver in Various Application of Consumer Products (Adopted from Lem et al., 2012) (January 1, 1980 to December 31, 2010)

High dose formulation applications were in antimicrobial pharmaceuticals, hair mousse, biocides, disinfectants, electronic chemicals, silver conductive ink, medical applications, solar panels, smart glass, and suppository (Lem et al., 2012).

4. The landscaping of consumer products containing nanosilver in Korea

As indicated earlier, Korea is a leader in patent publications for use of nanosilver in consumer products. Cosmetics and personal care, medical, and health care occupy 65% of the Korean applications shown in Figure 4. This is a similar percentage as the global consumer products shown in Figure 3. Comparing Figure 3 and Figure 4 illustrates some of the Asian cultural and societal differences. While the world is spending more (40%) on medical/ health care and little less (32%) on cosmetics/ personal care, Korea invested more than twice in cosmetic/personal care (47%) than in medical /health care (18%) (Lem et al., 2012).

Koreans spent about $130 per person in 2008 on makeup and skin care products and use in plastic surgery clinics doubled to about 1,000 facilities between 2004 and 2007 (Shin, 2008; Barry, 2002; U.S. Commercial Service, 2011; Consumer Demand Beneficiaries in Korea, 2008). Several reasons accounted for the strong growth of cosmetics in the Korean market.

1. Life expectancy of Koreans has increased from 62 in 1970 to 79 years in 2006 (Shin, 2008).
2. Korea's rapidly aging population and increasing female workforce are now driving the demand for beauty and personal cares products in the domestic market. Also the

life expectancy of female Korean has increased from 66 in 1970 to 82 in 1986 (Shin, 2008).

3. A notable trend is the rising demand of the male consumer. Male Korean life expectancy has increased from 59 in 1970 to 79 in 2006 (Shin, 2008).
4. The younger populace is looking for general skin care and hair care products while the older generation has more specific needs for their cosmetics products.
5. There is a clear trend of the market heading towards premium cosmetic products that need new technology, especially nanotechnology.

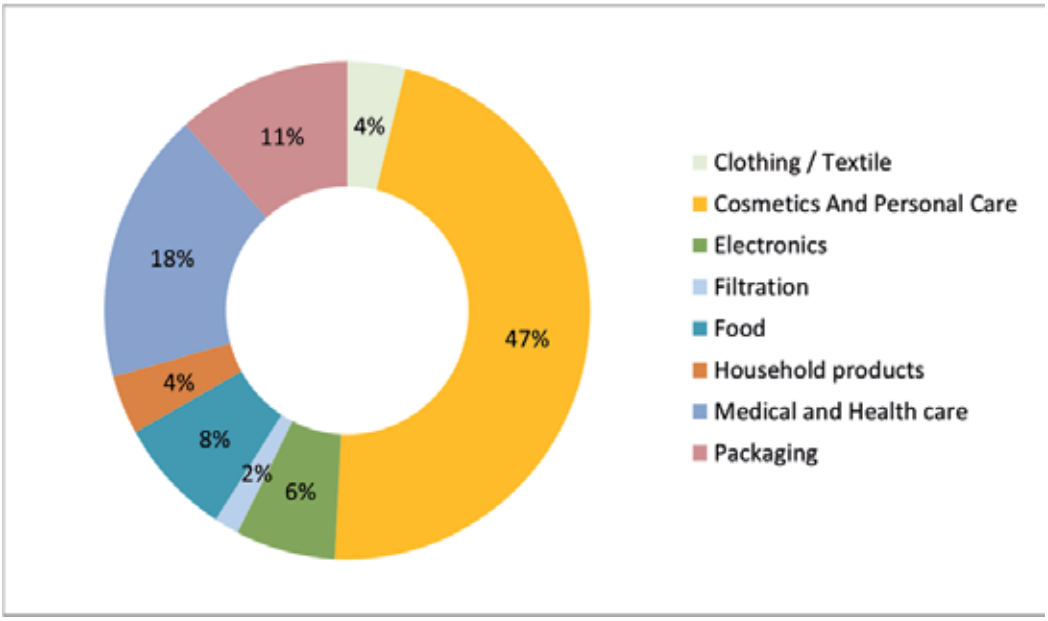


Fig. 4. Nanosilver in Different Application Areas in Korea (Adopted from Lem et al., 2012) (January 1, 1980 to December 31, 2010)

Recently, nanosilver has also found use in everyday products such as antimicrobial products, consumer products, and electronic products. Consumer products containing nanosilver have been selling everywhere in South Korea, especially on the streets of her capital – Seoul as illustrated in Figure 5.

At this juncture, it should be pointed out that not all the consumer products claimed to contain AgNPs indeed have AgNPs. Using energy dispersive X-ray spectroscopy (EDS) analysis (Yang et al., 2011), we have found that not all the commercial products sold in Korea claimed to have AgNPs indeed contain AgNPs (Kim et al., 2011; Yang et al., 2011). These experimental results are not unusual, though negative results can also be due to washing away over time. However, it appears that marketing (especially by street vendors) of AgNPs is not always directly tied to science. Even with a sophisticate Cloud Point Extraction-Based Separation together with EDS, SEM, TEM, and UV analyses, Chao et al., (2011) reported that only three out of six tested antibacterial AgNPs containing commercial products actually contained AgNPs.

Fig. 5. Selling AgNPs Consumer Products on the Street in Seoul, Korea

5. Silver/nanosilver and public health

Today, about 320 tons/year of AgNPs are produced and used worldwide in industrial products (Nowack et al., 2011). Stensberg et al. (2011) reported that an estimate of 1,120 tons of AgNPs will be used in 2015. They also reported that the number of products that contain AgNPs has increased from 30 in 2006 to over 300 at the beginning of 2011. In any event, the assumption that silver is benign to humans or that health effects are relatively mild cannot be made without further research to confirm these views. The existence of previous studies show, if anything, that there are potentially very real and severe side-effects which must be addressed prior to increasing the use of silver and especially nanosilver in health, consumer, and professional settings. Therefore, whether it is an old or new problem, we must deal with it seriously.

The growth in investment in nanoscience and nanotechnology has been astounding. Among the $12.4B spent in 2006 worldwide nanotechnology research funding (Mamikunian, 2007), at least 50 %, that is >$6B, was spent on the effect of size on the development of nanomaterials and devices.

Nanotechnology application focuses on exploitation of the size effects to create structures, devices and systems with novel properties and functions. The focus of nanoscience research is the understanding of the effect of size and its influence on the properties of nano-material.

With this rapid growth of nanotechnology, it is only natural that the next major wave of applications for silver would include nanoscale particles. However, the safety of nanosilver (AgNPs) in public health is a potential "Nano-Titanic" possibly preventing a sustainable nanosilver industry. As with macroscale silver, nanosilver effectively kills bacteria and is therefore biocidal, but many scientists are still uncertain of its safety.

At the nanoscale, materials have different properties as a function of size compared with the same materials at a larger size. The size range of greatest interest is typically from 100 nm down to approximately 0.2 nm, because in this size range properties of the materials become tunable (Sun, 2007; Lem et al., 2010). This tuneability requires a better understanding of effect of size of particles on public health.

Given the explosive growth in applications of nanosilver, certain authors have expressed misgivings about potential public health effects. Senjen and Illuminato (2007; 2009) of Friends of the Earth claimed that nanosilver was an extreme germ killer which presents a growing threat to public health. Chaloupka et al. (2010) discussed a number of medical uses of Ag and AgNPs in human prophylactic antibacterial effects such as bone cement, implants, and coating for neurosurgical shunts and catheters. In their literature citation, Wijnhoven et al. (2010) and Chao et al. (2010) found that AgNPs were toxic to rat and human cells. They further noted that these silver nanoparticles could enter the human skin via textile and dressing contact, via release from medical devices ingressing into the female gentile tract, forming protein-silver complexes that can deposit in human vital organs such as the liver, lungs, and kidneys. Recently, silver has found use in everyday products such as antimicrobial products, consumer products, and electronic products.

Even with the urgent need to specify the nanotechnology that could be the most assist the developing world, Faunce & Watal (2010) reported that the role of AgNPs was not specifically mentioned in water purification due to their potential environmental toxicity. Very recently, Powers (2010) mentioned in her dissertation that her results showed positive that Ag+ and AgNPs are developmental neurotoxicants in vitro and in vivo. Furthermore, the discharge, emission, and disposal of AgNPs and their products to the environment during their entire products life cycle are also a major concern (Rebitzera et al., 2004; Ross et al., 2002; Panyala et al., 2008; Hansen, 2009; Danscher & Locht, 2010; Nowack, 2010). In recent years, the uncertainty of safety has increasingly made nanosilver a concern of potential threats to public health. Despite the fact that silver and nanosilver has been used for many centuries in applications pertinent to our daily life because silver has an antiseptic effect. In ancient times, many Greeks used silver vessels for drinking water storage.

In contrast, Volpe (2010) and Height (2009) of The Silver Nanotechnology Working Group (SNWG) argued that AgNPs used in antimicrobial applications are identical to all the EPA-registered silver products that were used for decades. Very recently, Nowack et al. (2011) urged the policy regulators should not hastily declare nanosilver materials as new chemicals in their study on the 120+ years of nanosilver history. However, Schäfer et al. (2011) rebuked Nowack et al. (2011) by questioning the difference between the scientific definition of "colloid silver" and nanosilver. They further raised five pertinent questions regarding the safety of nanosilver in consumer products that need to be clarified:

1. Is the toxic potential of nanosilver identical to "classical" silver?
2. Since when has it been possible to analyze silver at the nanoscale?

3. Does nanosilver enter the body in the same way as "classical" silver?
4. Do we know enough about the environmental spread of silver resistance?
5. Is our current knowledge on nanosilver in consumer products sufficient to account for safe use?

In her exhaustive study, Powell (2011) clearly concluded that, "I propose that enough is known already about the toxicity of silver as a metal to begin taking strong steps to prevent human exposures and environmental releases now, rather than waiting till silver becomes the next mercury."

Recently, the German Federal Institute for Risk Assessment (BfR) has conducted the Delphi study (Bartels, 2010) regarding nanoscale silver compounds in food products, cosmetics and every day products. To ensure that products are safe for consumer health, BfR recently recommended that German manufacturers not use nanoscale silver or nanoscale silver compounds in foods and everyday products until the data are available and comprehensive enough to allow a conclusive risk assessment (Bartels, 2010). Faunce & Watal (2010) noted further the uncertainty of the safety may be compounded by lack of toxicological data and lifecycle studies of acceptable environmental exposure limits.

6. A clinical medicine perspective of silver and nanosilver

In view of the many applications in Asian countries like Korea, the importance of a clinical medicine perspective of silver and nanosilver needs to be emphasized. Silver has long been used as an antimicrobial agent in medicine to keep wounds clean since the days of ancient Greece, Egypt, and Rome (Chen & Schluesener, 2008; Lansdown, 2010). This is because of the efficacy of thiol group reactions which inactivate bacterial enzymatic activity (Faunce & Watal, 2010). Colloidal silver was introduced as long ago as 1884 by German physician Dr. C.S.F. Crede, to prevent transmission of maternal gonorrhea to newborns and thus preventing blindness. This is still a practice used in nurseries today (Feder, 2005). Silver was used as a wound dressing and disinfectant during World War I until the advent of penicillin, but the combination of silver with the antibiotic sulfonamide into silver sulfadiazine cream is still the first-line treatment for burns (Atiyeh et al., 2007; Ahamed et al., 2008; Faunce & Watal, 2010).

Nanosilver, or nano-particle sized silver provide a greater surface area of silver and theoretically a more efficacious product. It has also been used extensively in medical applications, from the impregnation and coating of surgical mesh, indwelling catheters, ports, stents, tubes, scopes, and cuffs to other devices to prevent the growth of bacterial biofilms which can precipitate infection. (Faunce & Watal, 2010) Additionally, the use of nanosilver has extended into the public health arena, where it is being used to coat food and agricultural facilities in an effort to prevent bacterial outbreaks in the general population (Powell, 2011).

Despite the wide ranging applications of silver and nanosilver in medicine, it is not clear that enough regulation exists or that there is sufficient research on the potential toxicity of silver to the human body and environment at large. Not only is there evidence for the potential toxicity of nanosilver, there is evidence to suggest that nanosilver is uniquely harmful to the human body when compared to silver compounds because of its ability to generate reactive oxygen species (ROS). ROS, also known as free radicals, cause a

biochemical chain reaction which eventually leads to the destruction of cellular metabolism, structures, and DNA (Faunce & Watal, 2010). Oxidative stress can disrupt cell membranes or cell walls, leading to cell destruction (Powell, 2011). In vitro studies have shown that silver can increase the rate of cell death, inhibit cell growth, decrease DNA and protein synthesis, disrupt DNA replication, affect cell membrane ion transport and integrity, cause cell swelling and toxicity, cause cell death, inhibit neutrophil and lymphocyte activity, and decrease the body's cell count (Powell, 2011).

Environmental concerns surrounding nanosilver have also entered the public health arena (Yu, 2008). Nanosilver production can lead to bulk form release of silver and nanosilver into waste streams, which have previously led to major environmental toxicities (Faunce & Watal, 2010). Such pollution can lead to not only deleterious effects to the ecosystem as a whole, but also cause direct poisoning of humans and animals. In fact, ionic silver is considered to be the second most toxic metal after mercury, in part because of its efficacy in binding prokaryotic and non-mammalian organisms (Power, 2011).

One common condition linked to increased silver deposits in the skin is argyria, a permanent blue-gray discoloration of the skin, and the related argyrosis, a similar discoloration in the eye (Atiyeh et al., 2007; Powell, 2011). The presence of silver is thought to increase melanin production, therefore leading to the color change, particularly in the presence of sun exposure (Powell, 2011). Though long thought to be a harmless cosmetic change, the finding of argyria is a proxy for increased systemic contamination with silver and suggests deeper pathophysiological effects in the body.

Medically, well-documented effects have suggested that silver harms the renal and hepatological systems (Powell, 2011). Deposits of silver in the glomerular subunits of the kidney have led to its classification as a nephrotoxin (Powell, 2011). In the cardiovascular system, case reports have noted an association with arteriosclerosis, a precursor of coronary artery disease (Powell, 2011). The inhalation of silver in the respiratory system has also been linked to inflammation, emphysema, reduction of lung volume, and straining of the tissues in the lung (Powell, 2011). As a result, patients have complained of sometimes daily upper respiratory tract irritation, cough, wheezing, and chest tightness (Powell, 2011). Silver has also been used as an abortifacient and sterilizing agent, suggesting its intrinsic damage to the reproductive tract (Powell, 2011). In pregnant women, silver has also been show to cross the placental barrier, allowing it to enter the fetus as well (Powell, 2011). One case-control study also suggested an association between the presence of silver in drinking water and developmental abnormalities (Powell, 2011).

Delayed wound healing and decreased white cell count has been found to be a result of silver-enhanced wound dressings (Powell, 2011). In patients with argyrosis, discoloration of the eyes with decreased night visual acuity has also been reported (Powell, 2011). Neurological effects include deposits in the central nervous system, glial changes, and cellular gliosis (Powell, 2011). Clinical manifestations may include seizures, vertigo, weakness, gait disturbance, and decreased sensation (Powell, 2011). Finally, studies in mouse embryonic stem cells have shown a rise in levels of p53, one of the main tumor suppressor proteins which help prevent cancer in the body, which inevitably leads to the question of whether nanosilver use can potentially contribute to greater likelihoods of cancer (Faunce & Watal, 2010). On the other hand, enhanced efficiency of wound healing has also been reported by application of AgNPs on skin wounds in mice (Liu X et al., 2010).

Novel research done recently by Powers (2010) has shown that monovalent silver impairs mechanisms of neuronal development in vitro, but also causes disruption of neurodevelopmental mechanisms in vivo, which persists as lasting changes in adult neurochemistry and behavior (Powers, 2010). Working with the model organism, zebrafish (Danio rerio), Powers was the first to show that while lower levels of silver ion did not affect morphology or embryonic viability, they do nevertheless negatively impact swimming performance and thus long-term mortality. Higher concentrations of ionic silver resulted in clear embryonic problems such as delayed hatching, decreased survival, and dysmorphology, suggesting a concentration-dependent effect (Powers, 2010). Powers also provides evidence for the teratogenic effect of silver nanoparticles and of note, shows that the toxicity of nanoparticles was through a distinct mechanism from ionic silver. Some of these biological effects can be explained through differing toxicokinetic and toxicodynamic effects, which elicited unique developmental and neurobehavioral pathologies (Powers, 2010).

Another study conducted by Seoul National University showed that nanosilver enhances platelet activation and procoagulant activity (Jun et al., 2009). Nanosilver worked synergistically with thrombin, a native blood protein which precipitates platelet activation and aggregation, to amplify thrombotic effects. Jun et al. (2009) found that nanosilver works in separate ways to enhance both the activation of platelets as well as facilitate platelet aggregation, or clumping. Intracellular calcium levels were increased by more than two fold in the presence of nanosilver, which is directly related to the activation of GPIIb/IIIa, a protein found on platelet surfaces which aids in platelet activation and binding to fibrinogen. By potentiating these platelet activation and aggregation pathways, nanosilver exposure can theoretically lead to increased thrombotic events. Thrombosis can lead to decreased blood flow or infarction in the circulatory system, particularly in individuals already predisposed to blood clots. Potential complications include venous thromboembolism, deep vein thrombosis, pulmonary embolism, stroke, and myocardial infarction, underscoring the importance of understanding the thrombotic effects of nanosilver (Jun et al., 2009)..

Despite the plethora of studies which suggest certain harmful effects of silver, the data are incomplete. The links between in vitro and animals studies to humans are in dispute, and the existing data on humans are largely through case reports, case-control studies, or retrospective analyses. There is a lack of high-quality, randomized controlled trials (RCTs) which can increase statistical power while minimizing biases. Obvious ethical concerns limit the amount and type research that can be done on human subjects, though longer-term, broader retrospective studies of patients exposed to silver may prove to be more helpful (Powell, 2011). In any event, the assumption that silver is benign to human begins or that health effects are relatively mild cannot be made without further research to confirm these views. The existence of previous studies show, if anything, that there are potentially very real and severe side-effects which must be addressed prior to increasing the use of silver and especially nanosilver in the health, consumer, and professional settings.

7. Concept of waste generation

Waste generation is a critical limitation for any sustainable industrial system (Evens et al., 2009). Waste is an important part of our life. Humans are not perfect and thus create wastes.

The actual performance is often lower than the theoretical because the efficiency is always less than 100% (Berglund and Snyder, 1990).

Elimination and minimization of waste have been making great progress in industries and businesses using advanced methodologies such as Lean Six Sigma (Curran et al., 2006; EPA, 2009). For instance, in polymer composite manufacturing, the total waste generated from these processes could be as high as 25% based on the theoretical yield of raw materials (Lem et al., 2006). Figure 6 gives a typical schematic of a materials processing value chain. The steps of the process include component selection, processing, structure, product, and performance. Each step can generate waste only if the waste cannot be recycle back to the start or inputs. Gutowski (2002) has examined the product induced material flows through the product manufacturing system, and has suggested several research strategies to reduce material related environmental loads. This can be accomplished by focusing on three key aspects of the manufacturing process: (a) resource productivity, (b) cleaning products, and (c) re-manufacturing, recycle, and compositing.

A simple mass balance of the net flow performance balance in Figure 6 can be described by Eqn 1.

$$\begin{aligned} Performance \;=\; & af\ (critical\ component\ selection\ -\ its\ wastes) + \\ & bf\ (process\ -\ its\ wastes) + cf\ (structure\ -\ its\ wastes) + \\ & df\ (products\ -\ its\ wastes)\ -\ performance\ wastes \end{aligned} \tag{1}$$

$$\begin{aligned} Performance \;=\; & af\ (critical\ component\ selection) + bf\ (process) + \\ & cf\ (structure) + df\ (products) - \sum (wastes)_{all\ sources} \end{aligned} \tag{2}$$

Where a, b, c, d are constants. In term of a continuous flow process, we can rewrite Eqns 1 and 2 into Eqn 3

$$P(x) = \int \lambda_j V(x_j) dx_j - \int \omega_j W(x_j) dx \tag{3}$$

Where, P(x) is a value performance function, $V(x_j)$ is the value generating function at component x_j stage or phase j, and $W(x_j)$ is the waste generating function at component x_j, and λ_j , and ω_j are constants.). The variation of V (x_j), $W(x_j)$, λ_j , and ω_j greatly affects the value of P(x).

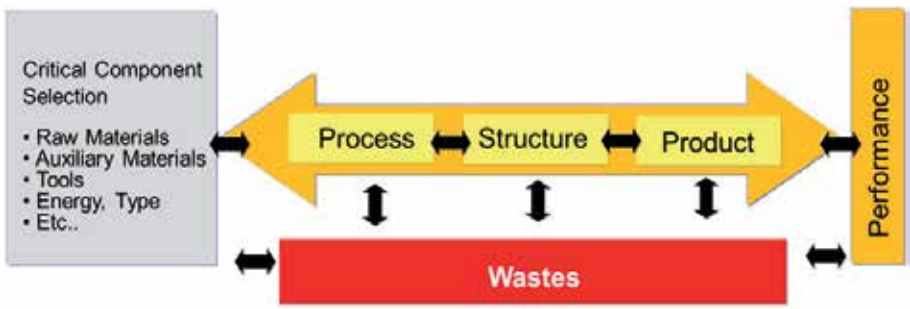


Fig. 6. Relationship of Component Selection–Processing-Structure-Product-Performance (Adopted from Lem et al., 2006)

The expression in Eqn 3 has found uses in many applications in science and engineering. An example of such is the tensile modulus of several ordered polymers (Lem et al., 2006) in Table 1, in which the actual value [P(x)] is substantially lower than the theoretical value [V(x)]. V (x_j) is not restricted with any limitations in Eqn 3 and it is valid to include the feedback loops (as in a recycling process) in Figure 6 except with different forms of V (x_j) and W (x_j), and different values of λ_j and ω_j.

#	Materials	Molecular Structure	Theoretical [V(x)] [value generating function at component x]	Actual [P(x)] [Value performance function]	Gap [W(x)] [Waste Generating Function]	
			GPa	GPa	GPa	%
1	Poly(p-phenylene-2,6-benzo[1,2-d:45-d'] bisoxazole (PBO)	Cis	730-670	360	370 – 310	55 – 43
		Trans	707-620		347 – 260	60 - 37
2	Poly(p-phenylene-2,6-benzo[1,2-d:45-d'] bisthiazole (PBZT)	Cis	610-600	325	285 – 275	48 – 45
		Trans	605-525		280 – 200	53 – 33
3	Polyethylene		360 – 320	172 – 117	243 – 148	76 – 41
4	Graphite		1500	600 – 70	1430 – 900	95 – 60

Table 1. Tensile Modulus of Several Ordered Polymers (Adapted from Lem et al., 2006)

8. Waste minimization in flow of materials as a food metabolism process in a material life cycle

The quality of our life is improved by our industrial system, but the current system is creating unintended and serious consequences for the environment and public health at a global scale. For nanotechnology, to minimize these consequences, one must be able to transform all sources of waste and toxicity into "technical" or "biological nutrients". We can then reuse them indefinitely without harm to living systems.

Senge and Carstedt (2001) offered a view of why industry produces waste and suggested that a synthetic process can emulate nature to reduce the waste using a cyclic industrial system. One example of this approach is recycling of nylon 6 carpets (Sifniades et al., 1999; Lem et al., 2001, 2002). This type of cyclic process has addressed and overcome the economic, technical, and logistical barriers to commercialize a closed loop recycling process and recover caprolactam from waste nylon 6 materials (Lem et al., 2010). Based on the exergy analysis by Dewulf et al. (2002-2007), in Figure 7, Lem et al. (2010; 2011) have shown that waste generation in a real process is more than just exergy loss (destroyed) in industrial metabolism. Waste generation is unavoidable so waste minimization becomes a fundamental requirement for economic feasibility. The energy and exergy concepts can be formulated in the laws of thermodynamics. Energy is motion or ability to produce motion. It is always conserved in a process (1st law). Exergy is work or ability to produce work. It is always conserved in a reversible process, but is always consumed in an irreversible process (2nd law, the law of exergy) (Wall, 1988; Sciubba & Wall, 2007))

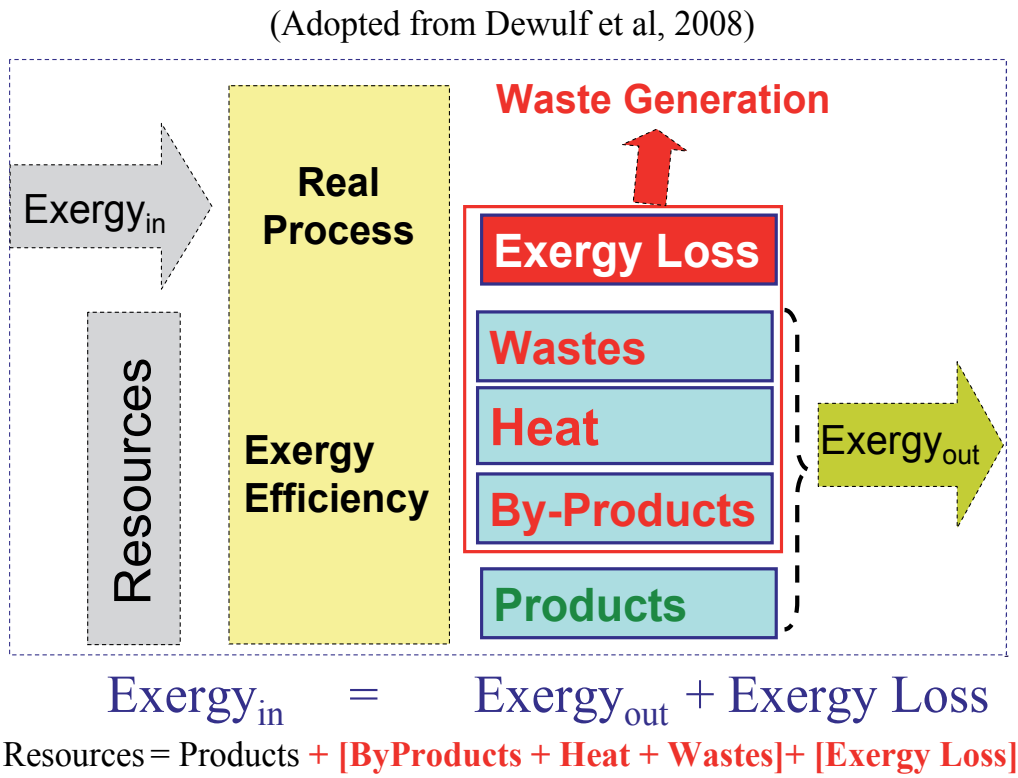


Fig. 7. Waste Generation and Exergy Loss (Dewulf et al., 2008, Lem et al., 2009, 2010)

9. Flow of Ag and AgNPs as a food metabolism process in material life cycle

Using a material flow analysis (MFA), Johnson et al. (2006) in their "anthropogenic cycling of silver in 1997" study have found that North America and Europe have the biggest share of use of silver products on a per capita basis. They found that global silver discards are approximately 57% of the silver mined and only 57% of the silver entering waste management globally is recycled. The amount of silver entering landfills globally is comparable to the amount found in silver mining tailings. Eckelman and Graedel (2007) reported that more than 13 Gg of silver are emitted annually to the environment globally. The tailings and landfills make up almost three-fourths of the total emission.

Figure 8 gives an overview of a silver/nanosilver product's life cycle as food and waste in industrial metabolism. The metabolization of resources should be optimized with respect to exergy. Dewulfn and Van Langenhove (2002, 2004) have previously applied exergy analysis as a quantitative tool in the thermodynamic optimization of the life cycle of plastics.

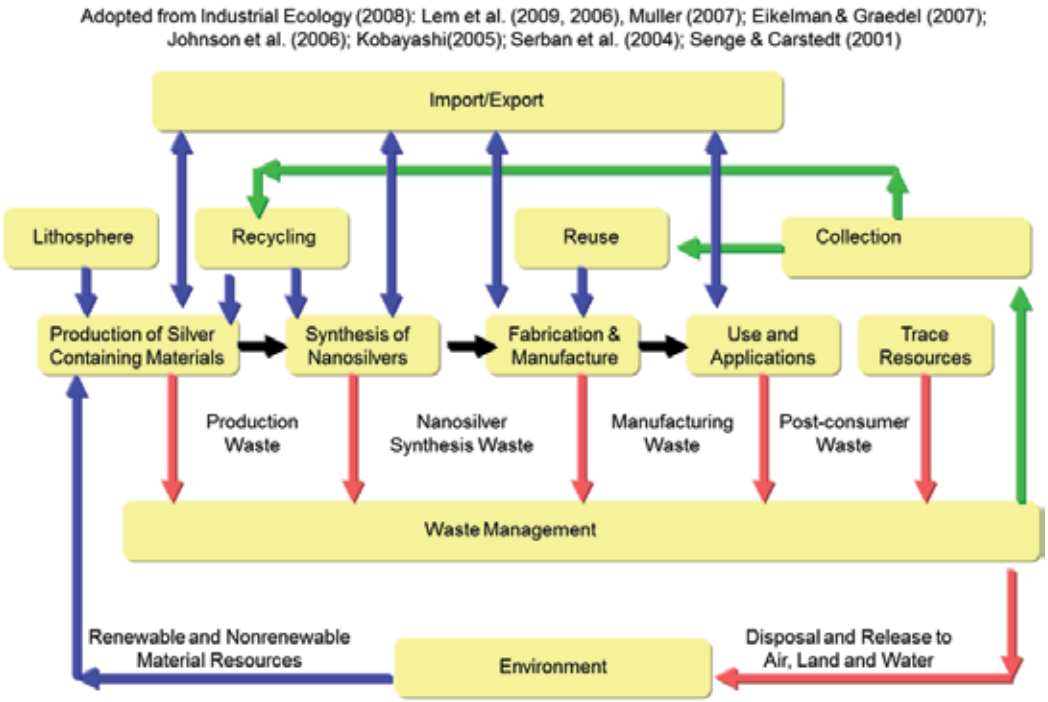


Fig. 8. A Silver/Nanosilver Product's Life Cycle as **Food** and **Waste** in Industrial Metabolism (Adopted from Lem et al., 2010)

Once again as in Figure 6, a mass balance of each step in the life cycle in Figure 8 is equal to the food resource (in blue color arrows) available in each step minus the wastes (in red color arrows) at each step. Therefore, a summation of all the steps gives rise to the total value generated. Eqn 4 can be found

$$Total\ Value\ Performance\ = \sum (Food\ Resource\ in\ Each\ Step)\ - \sum (wastes)_{all\ sources} \tag{4}$$

For the continuous process we have the generalized Eqn. 3 (above)

$$P(x) = \int \lambda_j V(x_j) dx_j - \int \omega_j W(x_j) dx \tag{5}$$

Therefore, the main thrust in the waste minimization is to minimize the waste generation function W (x_j) at any step j.

10. Effect of size on functional materials (silver)

It is well established that the size of nanomaterials affects its properties (Sun, 2007). There is no exception in AgNPs, particularly as an antibacterial and anti-biofouling agent (Chaloupka et al., 2010; Liu H-L et al., 2010; Liu JG et al., 2010; Sotiriou & Pratsinis, 2010). Fundamental morphology, surface area, and property changes with smaller size have led to

size dependent material properties which are substantially different from their counterparts in bulk. The extent of valence electron delocalization can vary with the size of the particle or domain. Quantum effects become relevant for sizes less than 10 nm. Material properties become tunable by size (Sun, 2007); notably, coordination number imperfection, surface relaxation behavior, nanosolidification in physical properties, superplasticity in mechanical properties, melting and thermal diffusivity in thermal properties, acoustic phonon hardening and optical phonon softening behavior, quantum confinement effects in optical properties, work function and dielectric suppression in electrical properties, and magnetic modulation in magnetic properties. For example, the bandgap of semiconductors such as ZnO, CdS, and Si, changes with size. Magnetic materials such as Fe, Co, Ni, Fe_3O_4, etc., exhibit size dependent magnetic memory properties (Sun, 2007).

In spite of the significance in the size of nanosilver, patenting directly addressing size effects only started in 2006 as seen in Figure 9. It is growing every year and 19 patent publications mentioned the size of nanosilver in 2010.

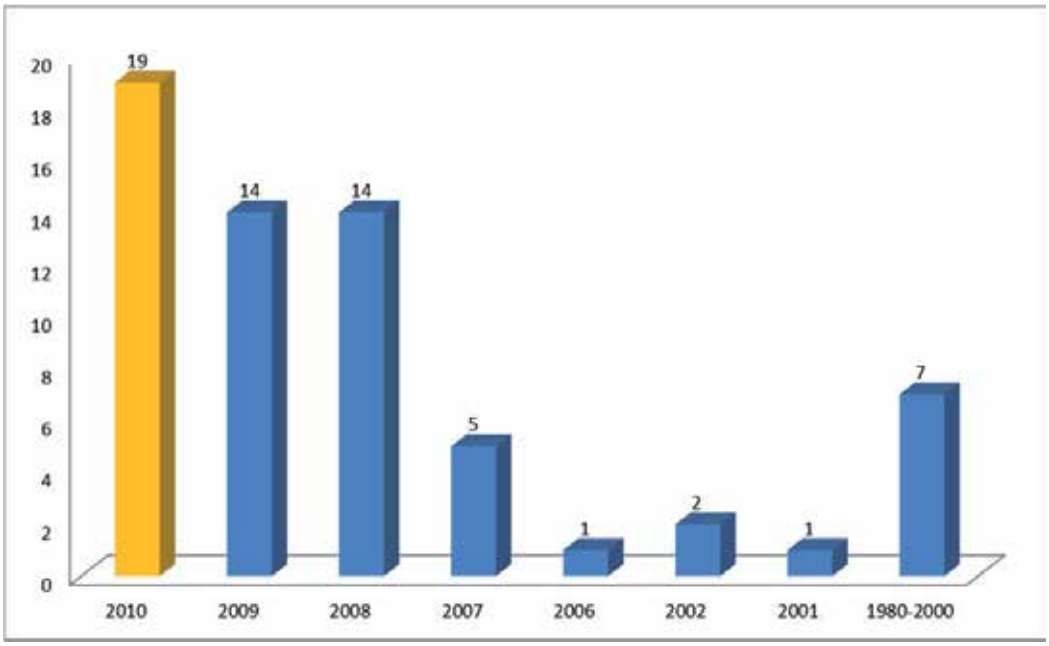


Fig. 9. Patents Describing Size of Nanosilver (Adopted from Lem et al., 2012)

As seen in Table 2, much effort has been employed to refine the type of stabilizers depending on the size of the nanosilver. For the larger diameters up to 400 nm, polyethylene glycol, poly (styrenesulfonate), cetyltrimethylammonium bromide have been used. For the medium diameters up to 100 nm, proteins, peptides, polyvinylpyrrolidone, human serum albumin and transferring have been reported in the IP publication to stabilize the nanosilver. For the very small diameter up to 15 nm, polyvinylpyrrolidone, (1-vinyl pyrrolidone)-acrylic acid copolymer, polyoxyethylene stearate, and 1-vinylpyrrolidone-vinyl acetic acid copolymer were used. Since most AgNPs require to be capped by stabilizers for dispersion, the cytotoxic effect from NP size may be mixed with that from stabilizers. A study employed physically produced AgNPs for examination of the size effect

(Liu H-L et al., 2010). Results revealed that AgNPs of smaller average size (among 3 nm, 6 nm or > 10 nm) had greater antibacterial activity as well as cytotoxicity. This study pointed out the critical role of NP size in their effect on human health and environment.

11. Waste minimization in eco-product design for public health

We have recommended earlier to use Design for Lean Six Sigma - Green (DFLSS-G) and TRIZ to design eco-products (Lem et al., 2009). Kobayashi (2005) has used a product life planning methodology based on a quality function deployment (QFD) and a software tool to establish an *eco*-design concept of a product and its life cycle in multigenerational eco-products development. Serban et al., (2004) have used a TRIZ approach to design for environment for over a product life cycle. We need to answer the following three hard questions in this design:

1. Do we have a complete understanding of AgNPs product life cycle?
2. Do we have a clear understanding on the unmet needs?
3. What can we do to minimize use of AgNPs with optimal effects?

Publication Number	Stabilizers/Important Components	Application Area/ End Product	Nanosilver Size
WO2010091529A1	Stabilizers: Proteins And/Or Peptides And/Or Polyvinylpyrrolidone: Human Serum Albumin And Transferrin;	Cosmetics And Personal Care / Hair Care	1-100 nm in diameter
US20100172997A1	Stabilizer : Agarose, Hydrogel, Paa (Poly Acrylic Acid), Pva (Poly Vinyl Alcohol), Chitosan, Pnipam (Poly-N-Isopropyl Acrylamide), Substituted Pnipam (Including Pnipam-Aa (Poly-N-Isopropyl Acrylamide-Acrylic Acid), Pnipam-Allylamine (Poly-N-Isopropyl Acrylamide-Allylamine), And Pnipam-Sh), Pamam (Polyamidoamine), Peg (Polyethylene Glycol), Alginic Acid and/or Hpc (Hydroxyl Propyl Cellulose)		
US20090326614A1	Stabilizer : Polyethylene Glycol (Peg), Poly(Styrenesulfonate), Cetyltrimethylammonium Bromide;		1-400 nm
CN101402757A	Stabilizer: Amine Light Stabilizer.	Packaging	
US20090011046A1	Stabilizer: Proteins And/Or Peptides: Human Serum Albumin or Transferrin		1-100 nm.
US20080248086A1	Stabilizer : Hydroquinone, Hydroquinone Monomethyl Ether, T-Butyl Paracresol And Hydroxy Methoxybenzophenone, A Pigment, Or A Beneficial Agent.	Medical - Implant	
KR20080083499A	Stabilizer: Polyvinylpyrrolidone, (1-Vinyl Pyrrolidone)-Acrylic Acid Copolymer, Polyoxyethylene Stearate, And 1-Vinylpyrrolidone-Vinyl Acetic Acid Copolymer.		1-15 nm
US20080181931A1	Stabilizer: Acrylic Acid, Polyacrylic Acid, Poly(Ethyleneimine), Polyvinylpyrrolidone		
KR20060026362A	Stabilizer: Glycerin, Polyethylene Glycol, Ethanol, Ethylene Glycol, Propylene Glycol, Sorbitan Fatty Acid Alkylester And Its Ethylene Oxide, Hydrogenated Caster Oil		
US20050013842A1	Stabilizer: Polyacrylic Acid (PAA), A Poly(Ethyleneimine) (PEI), A Poly(Vinylpyrrolidone) (PVP), A Copolymer of Acrylic Acid (AA) with a Vinylic Monomer, Acrylic Acid	Biomedical Device - Lense	

Table 2. Type of Stabilizers Used (Adopted from Lem et al., 2012)

We have started to answer the first question by examining each step of the material flow in a metabolism during the life cycle as discuss earlier. The value generated is equal to the food resource available in each step minus the wastes at each step (Lem et al., 2009). In the material flow model, we need to include probabilistic method as suggested by Gottschalk et al. (2010) that is commonly being used in Design for Six Sigma (DFSS, Curran et al., 2006). To answer the second and third question, we need to understand how the use of nanosilver can be minimized based on specific needs in release and apply the DFLSS and TRIZ to generate innovative ideas for the eco-products design. As an exercise, we will use a shoe pad as an example as illustrated as in Figure 10.

Fig. 10. Shoe Pads (Adopted from Lem et al., 2012)

The amount of AgNPs release depends on the mechanics of the release. To prevent and control these occurrences, it is necessary to use "right amount" of suitable biocides to control fowl and kill microbes. Using a TRIZ approach (Terninko et al., 1998; Rantanen & Dom., 2002) in Figure.11, such a concept is proposed to use water activity as a means to control the water content of $AgNO_3$ in the nanofibers where these nanofibers have a shell and core structure. In addition to the controlled release of AgNPs, the use of the nanofibers is to produce a very high contact angle surface to prevent water absorption on the surface (i.e., the Lotus Leaf Effect).

AgNPs can be controlled release at least seven in ways:

1. Particle size,
2. Particle surface modification,
3. Oxidant availability,
4. Media composition,
5. Structured release materials (such as multilayer shell and core structured nanofibers),
6. Release device structure,
7. Locality.

The release can be by one, combination of several, or a combination of all. The first four have been demonstrated by Liu JG et al. (2010) experimentally that the release of AgNPs can be tuned. To understand better the mechanic of the release, we will extend the work by Schiesser (1992, 2011) to describe the release control (desorption and diffusion) in our model.

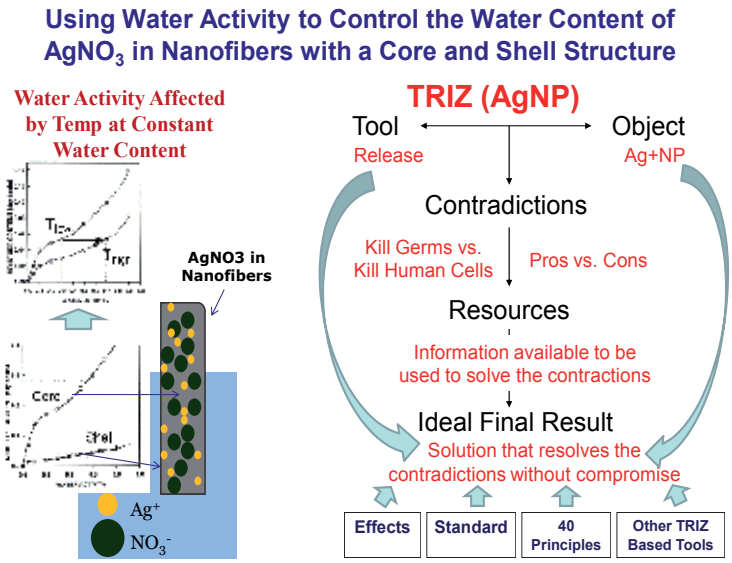


Fig. 11. The Proposed TRIZ Concept (Adopted from Lem et al., 2011)

11.1 DFLSS and TRIZ

A flow chart of the procedure to be used in our study is given in Figure 12 and a TRIZ approach in Design for Lean Six Sigma – Green for AgNPs products life cycle is given in Table 3. We are using the following four steps iterative approach:

First: determine the Voice of the Environment regarding the safety of the AgNPs products using two extreme sides of the debate between Friends of Earth/USEPA and Silver Nanotechnology Work Group (SNWG) to obtain a resolution regarding "Conflict". We try to answer the question - could improving one technical characteristic to solve a problem cause other technical characteristics to worsen? Once the problem is defined, we need to define the system boundaries, quantify mass flows of AgNPs, and define several emission scenarios.

Second: search for previously well-solved problems by looking at the 39 engineering parameters/40 principles (Terninko et al., 1998; Rantanen & Dom., 2002). Antimicrobial nanoscale silver is typically embedded within substrates, mainly a a matrix such as a polymer, where any antimicrobial functionality is achieved via release of silver ions (Ag+).

The behavior of silver in environment will be reviewed, and a mass balance model applied to calculate predicted environmental concentrations. The uncertainty of the results is assessed and predicted concentrations are compared to experimental and empirical data (examine an example such as "Nanoparticle Silver Released into Water from Commercially Available Sock Fabrics" by Benn and Westerhoff, 2008).

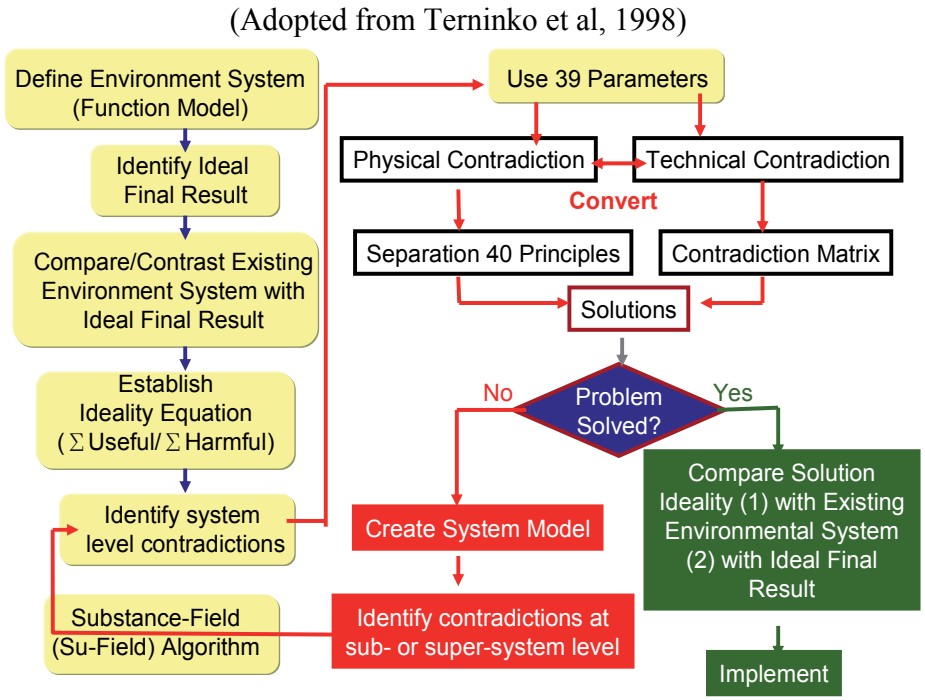


Fig. 12. Flow Chart for DFLSS-G with TRIZ (Adopted from Lem et al., 2010)

DFSS Phase	TRIZ Tools	Approach	Application to AgNP Product Life Cycle
Voice of the Customer	**1. Conflict Resolution, 2. Ideal Final Result, 3. Development of Measurement Systems.**	Identify the Problem	Step 1: Voice of the Environment (VOE) 1. Safety of AgNP Products. 2. Define Ideality Based QFD
Concept Development	**All**	1. Find The Principle that Needs to be Changed 2. Then Find the Principle that is an Undesired Secondary Effect.	Step 2: Conflict resolution 1. Example - Friends of Earth/USEPA vs. Silver Nanotechnology Work Group 2. Define Functionality/ Requirements
Detailed Design	**All**	1. Find the Principle that Needs to be Changed, 2. Then Find the Principle that is an Undesired Secondary Effect.	3. Use of Resources 4. Search for Previously Well-Solved Problems a. Examine 39 engineering parameters/40 principles. b. IP Landscaping
Optimize	**1. Conflict Resolution, 2. Trimming, 3. Subversion Analysis, 4. Problem Solving**	1. Look for Analogous Solutions 2. Adapt to the Potential Solution 3. Optimize – Ideality	Step 3: Review toxicity data for environmentally relevant silver compounds. Optimize wherever possible. Review earlier search for previously well-solved problems
Validate/ Implement	**1. Conflict Resolution, 2. Trimming, 3. Problem Solving**	Validate potential solution	Step 4: Gap Closing - Conflict resolution/Ideality Revisit

Table 3. TRIZ Approach in DFLSS-G for AgNPs Products Life Cycle (Adopted from Lem et al., 2010)

Third: compile and predict the toxicity data for environmentally relevant silver compounds for no effect concentrations. This material flow will be optimized based on a review of our earlier search for previously well-solved problems.

Fourth: evaluate and determine the potential for risk caused by the release of silver into environment using all available experimental data and literature data.

11.2 Release mechanics of AgNPs

As discussed earlier, the release of AgNPs can be controlled seven ways: (1) particle size, (2) particle surface modification, (3) oxidant availability, (4) media composition, (5) structured release materials, (6) structure of release device, and (7) locality. The release can be by one, combination of several, or a combination of all. The first four have been demonstrated by Liu JG et al. (2010) that the release of AgNPs can be tuned. The readers are referred to their excellent paper for details. In this section, we will focus our discussion on the last three methods.

11.2.1 Structured release material

One way to control the release of AgNPs is the control of the presence of water. Water activity (a_w) is defined as $a_w=p/p_o$ (where p and p_o are the partial pressures of water above a medium such as a food and a pure solution under identical conditions It is a measure of how efficiently the "free" water vs. the "bound" water present can take part in a chemical and/or physical reaction. Water content as a function of water activity has played a critical role in the understanding of food processing science and technology (Cassini et al., 2009). Nadia et al. (2011) have suggested further use of the glass transition temperature (Tg) of the material together with water activity in the material. This combination is a powerful tool for understanding the quantification of water mobility in foods and controlling the shelf-life of products. They reported that Tg, moisture content, and a_w are useful tools to quantify the water migration pattern in food precisely (Nadia et al., 2011).

The design of the structured release material must have an appropriate Tg, and the desired concentration of total water content present in a medium strongly bound to specific sites. These sites can be the hydroxyl groups of polysaccharides, the carbonyl, amino groups of proteins or synthetic polymers like nylon, polyurethanes, and other polar polymers containing hydrogen bonds and ion-dipole bonds. The preferred structure of the release material can be either bilayer such as shell/core or multilayered where the availability of free water in the material containing $AgNO_3$ can be controlled as needed (see Figure 13).

11.2.2 Structure of the release device

It has been known for many centuries that water forms spherical droplets on a leaf as seen in Figure 14, and it is more pronounced in the lotus leaf (Luzinov et al., 2006; Ramaratnam et al., 2008; Schilthuizen, 2009; Eichhoff, 2011). Lotus leaves are unusually water-repellent and keep themselves spotless, because on their surface there are countless miniature protrusions, coated with a water-repellant hydrophobic substance. Water cannot spread out on the leaves; so it acts as droplets, removing grime and soil as it moves. The rough surface inhibits wettability and reduces the contact area for dirt particles. Lotus effect has found many interesting applications in consumer products, surface coatings, electronic materials, and smart textile (Luzinov et al., 2006; Ramaratnam et al., 2008; Schilthuizen, 2009).

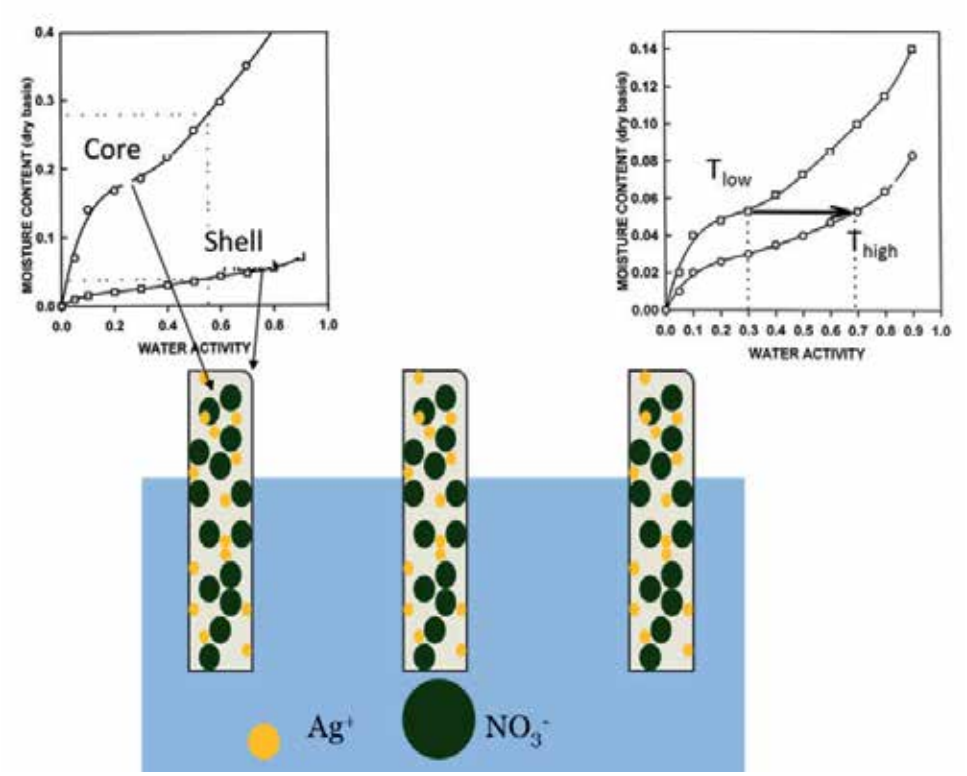


Fig. 13. The Proposed Structural Release Medium (Adopted from Lem et al., 2011)

Fig. 14. Water Droplets on a Leaf

Our goal is to control the wetting of water on the release device by using the concept advanced by Nano-Tex, LLC. Nano-Tex improves the water-repellent property of fabric using the so-called "Lotus Effect" by creating hydrocarbon nano-whiskers that are of 1/1000 of the size of a typical cotton fiber. The distance between the whiskers on the fabric is smaller than a typical drop of water and water thus remains on the top of the whiskers and above the surface of the fabric. (Eichhoff, 2011; Schneider, 2008; Wong et al., 2006; Lo, 2006).

A pictorial diagram of our proposed structure of the release device is shown in Figure 15. The materials used to make the release device have been suggested by KnollTextile (2010) and Wong et al. (2006).

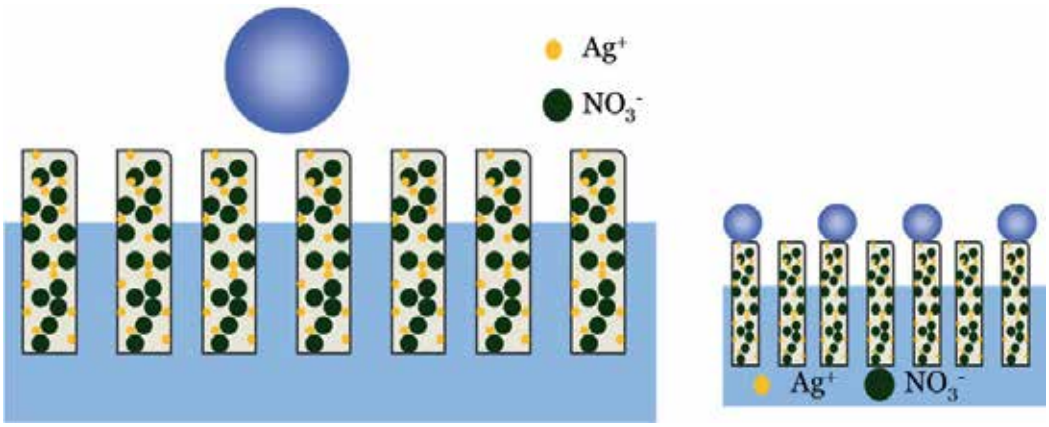


Fig. 15. Proposed Structure of Release Device (Adopted from Lem et al., 2011)

11.2.3 Locality

Bacterial fouling by humans has become a serious environmental and health issue. The existence of bacteria and its fouling in shoes and socks used/worn by human can lead to problems such as biofouling accumulation which leads to health problems. However, as seen in Figure 16, only certain areas in a shoe pad may require suitable biocides such as AgNPs for antifouling. Most sweat and frictional force occur in these areas indicated by the changing of the color of the pad.

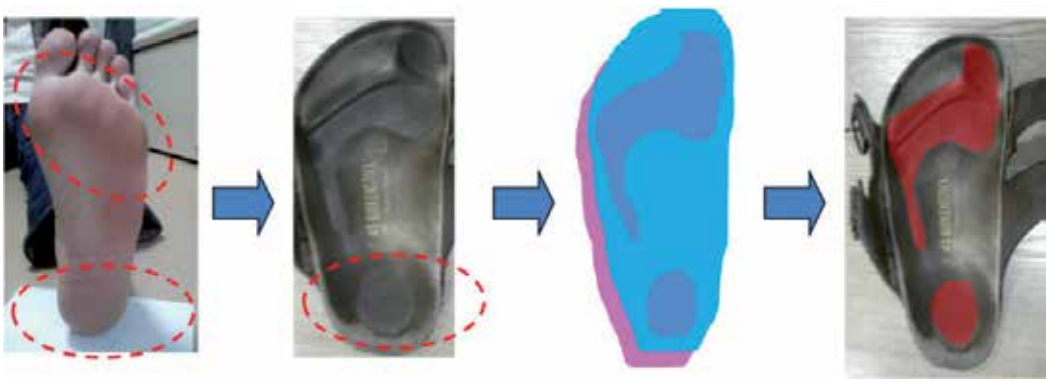


Fig. 16. Locality of Required Biocides for a Foot and a Shoe (Adopted from Yang et al., 2010)

12. Future study

In our Design for Lean Six Sigma based Waste Minimization research program, we have begun our journey to study the life cycle assessment of nanosilver starting with the use of product life cycle process mapping and Design for Lean Six Sigma with TRIZ. We are planning to have a more multidisciplinary and international interaction to the characterization of AgNPs products and their transformations in relevant biological and environmental media. A rigorous material flow analysis is needed to quantitatively assess the environmental impact of AgNPs emission. We have continued our study on waste minimization for the safe use of nanosilver in consumer products with particular attention paid to the eco-product design for public health. The data that have been generated from an IP search study help us design eco-products using Design for Lean Six Sigma - Green (DFLSS - G) and TRIZ (Curran et al., 2006; Lem et al., 2006; Terninko et al., 1998) as seen in Figure 17. In addition we need to verify the concept illustrated in Figure 17 experimentally. We will use Monte Carlo (Curran et al., 2006), artificial neural network modeling (Chayjan et al., 2011), and generic programming (Langdon, 2008) approach in the front-end of the innovative concept generation process to search for the best new generation design. To have a better understanding the mechanic of the release control, we will extend the work by Schiesser and his coworkers (Silebi & Schiesser, 1992; 2011) to describe the desorption and diffusion in a pore with Monte Carlo simulations (Gottschalk et al., 2010).

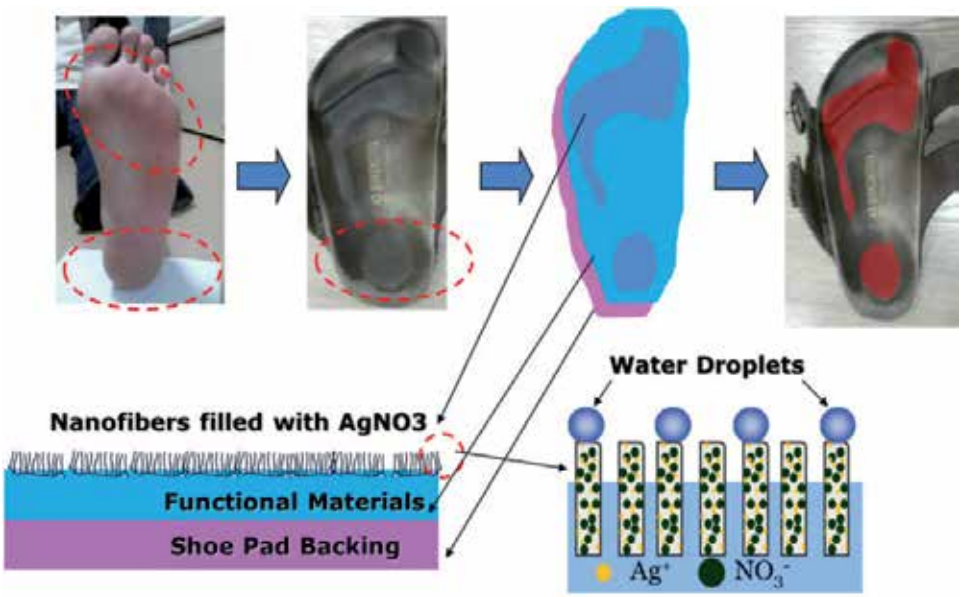


Fig. 17. Proposed Structure of an Eco-Product Required Biocides for a Shoe Pad (Adopted from Lem et al., 2011)

13. References

Ahamed, M., Karns, M., Goodson, M., Rowe, J., Hussain, S. M., Schlager, J. J., & Hong, Y.L., (2008), "DNA Damage Response to Different Surface Chemistry of Silver Nanoparticles in Mammalian Cells," Toxicology and Applied Pharmacology, 233, pp. 404–410.

Atiyeh, B. S., Costagliola, M., Hayek, S. N., & Dibo, S. A., (2007), "Effect of Silver on Burn Wound Infection Control and Healing: Review of the Literature", Burns, 33, pp. 139–148.

Barry D., (2002), "How to Win Face in the Korean Cosmetics Market, EXPORT AMERICA", December 2002, pp. 6-7.

Bartels K., (2010), BfR Opinion Nr. 024/2010. "Status of Regulation for Nanomaterials including Nanosilver in the EU in General and for Use as Biocides and in Novel Foods," IPINTECH LLC, November 30, 2010, Private Report to Nanobiz LLC. 28 December 2009; Available at Http://Www.Bfr.Bund.De/Cm/216/Bfr_Raet_Von_Nanosilber_In_Lebensmitteln_Und_Produkten_Des_Taeglichen_Bedarfs_Ab.Pdf

Benn, T.M., & Westerhoff, P., (2008), "Nanoparticle Silver Released into Water from Commercially Available Sock Fabrics", Environ. Sci. Technol., 42 (11), pp. 4133–4139

Berglund, R. L. & Snyder, G. E., (1990), "Minimize Waste during Design", Hydrocarbon Processing, International Edition. 69(4), pp. 39-42.

Brauer S., Lem K.W., & Haw J.R., (2009), "The Markets for Soft Nanomaterials: Cosmetics and Pharmaceuticals", Nano and Green Technology Conference. New York City, November 18, 2009.

Brumlik, C. J., Lem, K. W., Choudhury, A., Lakhani, A. A., Kuyate, P., Pathak, P. P., Vaidya, M., Iqbal, Z., & Careil, J-M., (2011), "Overview of 2010-2011 technology trends in nano-engineered energy generation and storage for large commercial markets," To be Presented at Nanotechnology 2011 Conference, Nanomaterials and Nanochemistry, Nano-Enabled Energy Systems, Nanomedicine and Nano-Bio Convergence - Emphasizing Emerging Science and Technologies, Applications, Commercialization and Business Opportunities, Javits Convention Center, New York, NY, November 1-3, 2011.

Cassini, A.S., Marczak, L.D.F., & Noreña, C.P.Z., (2009), "Comparison between the Isotherms of Two Commercial Types of Textured Soy Protein", Latin American Applied Research, 39, pp. 91-97.

Chaloupka K., Malam Y.K., & Seifalian A.M., (2010), "Nanosilver as a New Generation of Nanoproduct in Biomedical Applications", Trends in Biotechnology, 28(11), pp. 580-588.

Chao J.B., Liu J.F., Yu S.J., Feng Y.D., Tan Z.Q., Liu R, & Yin Y.G., (2011), "Speciation Analysis of Silver Nanoparticles and Silver Ions in Antibacterial Products and Environmental Waters Via Cloud Point Extraction-Based Separation", Anal.Chem., 83, pp. 6875-6882.

Chayjan, R. A. & Esna –Ashar, M., (2011), "Effect of Moisture Content on Thermodynamic Characteristics of Grape: Mathematical and Artificial Neural Network Modeling", Czech J. Food Sci., 29 (3) pp. 250–259.

Chen, X., & Schluesener, H. J. , (2008), "Nanosilver: A Nanoproduct in Medical Application", Toxicology Letters, 176, pp. 1–12.

Chih, Y-W., & Cheng, W-T., (2007), "Supercritical Carbon Dioxide-Assisted Synthesis of Silver Nano-Particles in Polyol Process", Materials Science and Engineering B 145, pp. 67–75.

Ciantar, C., Hadfield, M. & Howarth, G., (2001), "Case Studies to Assist Integrating Waste Prevention in Product Design", Meche Conference Transactions; Engineering for Profit from Waste, 9, pp. 201-210.

Cochrane, T., & Smith, J. A., (2001), "Designing Processes and Products to Minimize Wastes Produced", Meche Conference Transactions; Engineering For Profit from Waste, 9, pp. 137-148.

Consumer Demand Beneficiaries in Korea - Cosmetics Market, July 2008. Available At Http://Www.Invescogreatwall.Com/Data/20080821090357MI-Korea-Jul08-E.Pdf

Crosera, M., Bovenzi, M., Maina, G., Adami, G., Zanette, C., Florio, C., & Larese, F. F., (2009), "Nanoparticle Dermal Absorption and Toxicity: A Review of the Literature", Int Arch Occup Environ Health, 82, pp. 1043–1055.

Curran, S.A., Lem, K. W., Sund, S., & Gabriel,G., (2006), "Six Sigma Design: An Overview of Design for Six Sigma (DFSS)", Encyclopedia of Chemical Processing (Lee, S., Ed.), DOI: 10.1081/E-ECHP-120016185, Marcel Dekker, pp. 2719-2733

Danscher, G., Locht, L. J., (2010), "In Vivo Liberation of Silver Ions from Metallic Silver Surfaces", Histochem Cell Biol., 133, pp. 359–366.

De Meester, B., Dewulf, J., Verbeke,S., Janssens, A., & Van Langenhovea, H., (2009), "Exergetic Life-Cycle Assessment (ELCA) for Resource Consumption Evaluation in the Built Environment", Building and Environment, 44, pp. 11–17.

Deng K.E, & Jin M. Z., (2003), "Process For Preparing Colloidal Silver Solution", CN1433776A, 2003.

Dewulf, J., & Van Langenhove, H., (2002), "Assessment of the Sustainability of Technology by Means of a Thermodynamically Based Life Cycle Analysis", Environ Sci & Pollut Res., 9 (4), pp. 267-273

Dewulf, J., & Van Langenhove, H., (2004), "Thermodynamic Optimization of the Life Cycle of Plastics by Exergy Analysis", Int. J. Energy Res. 28, pp. 969–976

Dewulf, J., Van Langenhove, H., Muys, B., Bruers, S., Bakshi, B. R., Grubb, G. F., Paulus, D. M., & Sciubba. E., (2008), "Exergy: Its Potential and Limitations in Environmental Science and Technology", Environmental Science & Technology, 42(7), 2221-2232

Dewulf, J., Bössh, J.M.E., Demeester, B., Van Dervorst , G., H. Van Langenhove, H., Hellweg, S., & Huijbregts, M. A. J., (2007), "Cumulative Exergy Extraction from the Natural Environment (CEENE): a Comprehensive Life Cycle Impact Assessment Method for Resource Accounting", Environ. Sci. Technol., 41, 8477–8483

Eckelman, M. J., & Graedel, T.E., (2007), "Silver Emissions and their Environmental Impacts: A Multilevel Assessment", Environ. Sci. Technol., 41, pp 6283-6289

Edgar, T. F., & Huang, Y. L., (1994), "Artificial Intelligence Approach To Synthesis Of A Process For Waste Minimization", In Emerging Technologies In Hazardous Waste Management IV, ACS Symposium Series. 554, pp. 96-113.

EI-Badawy A, Feldhake D, & Venkatapathy R., (2010) "State of the Science Literature Review: Everything Nanosilver and More", Scientific, Technical, Research, Engineering and Modelling Support Final Report, United States Environmental Protection Agency, July 15, 2010.

Eichhoff, J. (2011), "Smart Textiles Creating Added Value For Textile Products," Friedrichshafen, 16 July 2011 Available At Http://Www.Outdoor-Show.De/Od-Wassets/Daten/Rahmenprogramm/Pdf/Smart-Textiles.Pdf

Evans, S., Bergendahl, M. N., Gregory, M., & Ryan, C., (2009), "Towards a sustainable industrial system - With recommendations for education, research, industry and policy," University of Cambridge Institute for Manufacturing, 2009.

Faunce T., & Watal A., (2010), "Nanosilver and Global Public Health: International Regulatory Issues", Nanomedicine 5(4), pp. 617-632.

Feder, B., (2005) "Old Curative Gets New Life at Tiny Scale." The New York Times, December 20, 2005

Gottschalk, F., Scholz, R. W., & Nowack, B., "Probabilistic Material Flow Modeling for Assessing the Environmental Exposure to Compounds: Methodology and an Application to Engineered Nano-Tio2 Particles", Environmental Modelling & Software, 25, 320, 2010

Gutowski, T. G., (2002), "Environmentally Benign Manufacturing and Ecomaterials; Product Induced Material Flows", Materials Transactions. 43(3), pp. 359-363.

Hansen, S. F., (2009) "Regulation and Risk Assessment of Nanomaterials – Too Little, Too Late?", PhD Thesis, Department of Environmental Engineering, Technical University of Denmark

Height M.J., (2009), "Evaluation Of Hazard And Exposure Associated With Nanosilver And Other Nanometal Oxide Pesticide Products," FIFRA Scientific Advisory Panel (SAP) Open Consultation Meeting. Arlington, Virginia, November 3 – 6, 2009.

Housenger, J. E., (2009) "Status of Regulating Nanoscale Particles and Prions," CPDA Mid-Year Meeting, March 11, 2009

Johnson, J. J., Jirikowic, J., Bertram, M., van Beers, D., Gordon, R. B., Henderson, K., Klee, R. J., Lanzano, T., Lifset, R., Oetjen, L., & Graedel, T. E., (2005), "Contemporary anthropogenic silver cycle: A multilevel Analysis", Environmental Science & Technology, 39, pp. 4655-4665.

Johnson, J., Bertram, M., Henderson, K., Jirikowic. J., & Graedel, T.E., (2005), "The Contemporary Asian Silver Cycle: 1-year Stocks and Flows", J Mater Cycles Waste Manag., 7, pp. 93–103

Johnson, J., Gordon, R., & Graedel, T.E., (2006), "Silver Cycles: The Stocks and Flows Project, Part 3", JOM, 2006 February, pp. 34-38

Jun, E-H, Lim, K-M, Kim, K. Y., Bae, O-N, Noh, J-Y., Chung, K-H., & Chung, J-H., (2009), "Silver nanoparticles enhance thrombus formation through increased platelet aggregation and procoagulant activity." College of Pharmacy, Seoul National University, Seoul, 18.

Karthik Ramaratnam, K., Iyer, S. K., Kinnan, M. K., Chumanov, G., Brown, P. J., & Luzinov, I., (2008), "Ultrahydrophobic Textiles Using Nanoparticles: Lotus Approach", Journal Of Engineered Fibers And Fabrics, 3(4), pp. 1-14 Available at http://www.jeffjournal.org/papers/Volume3/3.4.1_Brown.pdf

Kim, S-J., Kim, M. N., Jang, M.H., & Hwang, Y. Y., (2010), 'Kim Chi Containers Containing Nanosilvers," Fall 2010 Senior Course (KU 3176) Project, Department of Materials Chemistry and Engineering, Konkuk University, Seoul, Korea.

Knolltextile, (2010), "Nano-Tex® With Bioam Antimicrobial", January 2010 Available At Http://Www.Knoll.Com/Techdoc/KT_Tech_Bioam.Pdf

Kobayashi, H., (2005), "Strategic Evolution of Eco-Products: A Product Life Cycle Planning Methodology", Research in Engineering Design, 16 (1-2), pp. 1-16

Koecher J., Eiden, S., Mayer-Bartschmid, A., & Knezevic, I., (2009), "Medical Devices with an Antibacterial Polyurethaneurea Coating", US20090252804A1, 2009.

Langdon, W. B., (2008), "Genetic Programming for Drug Discovery", Technical Report CES-481 ISSN: 1744-8050, 26 February 2008.
Available at http://www.cs.ucl.ac.uk/staff/ucacbbl/WBL_papers.html

Lansdown, A. B. G., (2010), "A Pharmacological and Toxicological Profile of Silver as an Antimicrobial Agent in Medical Devices,", Advances in Pharmacological Sciences, Volume 2010, Article ID 910686, 16 pages

Lem K.W., Choudhury A., Lakhani A.A., Kuyate P., Haw J.R., Lee D.S., Iqbal Z., & Brumlik, C. J., (2012), "Use of Nanosilver in Consumer Products," Recent Patents On Nanotechnology, 6 (In Press)

Lem K.W., Haw J.R., Lee D.S., Iqbal Z., Salama A., Semthil Kurmaran, R., Sund, S., Curran, S., Brumlik, C., & Choudhury, A., (2011), " Nanosilver – Why It Is Still So Hot?", NSTI-Nanotech 2011, 3(7), pp. 557 - 560.

Lem K.W., Haw J.R., Sund S., Curran S.A., Brumlik C., Smith P., Brauer, S., Schmidt, D., & Iqbal, Z., (2009), "Waste Minimization in Commercialization of Nanotechnology", Seminar Presented Chonbuk National University, Jeonju, Jeollabuk-Do, South Korea, November 26, 2009.

Lem, K. W., Curran, S.A., Sund, S., & Gabriel, G., (2006) "Thermosets: Materials, Processes, and Waste Minimization", Encyclopedia of Chemical Processing (Lee, S., Ed.), Marcel Dekker, pp. 3031-3047, DOI: 10.1081/E-ECHP-120007991

Lem, K. W., Letton, A., Izod, T. P. J., Lupton, F. S., & Bedwell, W. B., (2001), "Composition Containing Caprolactam-Free Residue from Depolymerization of Nylon 6 Carpet and Use Thereof In Paving Asphalt, Plastic Lumber and Crack Sealants", US Patent 6,214,908, April, 10, 2001.

Lem, K. W., Letton, A., Izod, T. P. J., Lupton, F. S., & Bedwell, W. B., (2002), "Composition Containing Caprolactam-Free Residue from Depolymerization of Nylon 6 Carpet and Use Thereof In Paving Asphalt, Plastic Lumber and Crack Sealants", USP 6,414,066, July 2, 2002.

Lem, K.W., Haw, J. R., Lee, D.S., Iqbal, Z., A. Salama, Kurmaran, S., Sund, S., Curran, S., Brumlik, C., & Choudhury, A., (2010), "Waste Minimization in Consumer Products Containing Nanosilver" Seminar presented Chonbuk National University, Jeonju, Korea, December 2, 2010.

Lem, K.W., Haw, J. R., Lee, D.S., Iqbal, Z., A. Salama, Kurmaran, S., Sund, S., Curran, S., Brumlik, C., & Choudhury, A., (2011) "Nano Silver-Why It is still so Hot Now?", Environment, Health & Safety, NSTI-Nanotech 2011, 3(7), pp. 557

Lem, K.W., Haw, J. R., Lee, D.S., Brumlik, C., Sund, S., Curran, S., Smith, P., Brauer, S., Schmidt, D., & Iqbal, Z., (2009), "Waste Minimization in Commercialization of Nanotechnology" Seminar presented Chonbuk National University, Jeonju, Jeollabuk-do, South Korea, November 26, 2009.

Lem, K.W., Haw, J. R., Lee, D.S., Brumlik, C., Sund, S., Curran, S., Smith, P., Brauer, S., Schmidt, D., & Iqbal, Z., (2010) "Nano Silver - Why It is so Hot Now?", Nanoparticle Synthesis & Applications, NSTI-Nanotech 2010, 1 (3), pp. 391

Lem, K.W., Haw, J. R., Lee, D.S., Brumlik, C., Sund, S., Curran, S., Smith, P., Brauer, S., Schmidt, D., & Iqbal, Z., (2010), "Effect of Size on Properties of Nano-Structured Polymers - Transition from Macroscaling to Nanoscaling", Polymer Nanotechnology (Proceeding), NSTI-Nanotech 2010, 1(6), pp. 889

Liu J.G., Sonshine D.A., Shervani S., & Hurt R.H., (2010), "Controlled Release of Biologically Active Silver from Nanosilver Surfaces," ACS Nano, 4(11), pp. 6903-69013.

Liu, H-L., Dai, S. H. A., Fu, K-Y & Hsu, S-H., (2010), "Antibacterial Properties of Silver Nanoparticles in three Different Sizes and Their Nanocomposites with a New Waterborne Polyurethane", Int. J. Nanomedicine, 5, pp 1017-2028.

Liu, X., Lee, P.Y., Ho, C.M., Lui, V.C., Chen, Y., Che, C.M., Tam, P.K., Wong, K. K., (2010), "Silver Nanoparticles Mediate Differential Responses in Keratinocytes and Fibroblasts during Skin Wound Healing". Chem. Med. Chem., 5(3), pp.468-475

Lo, L. Y., (2006), "Wrinkle-Resistant Finishes On Cotton Fabric Using Nanotechnology," Ph.D. Thesis, Institute of Textiles and Clothing, The Hong Kong Polytechnic University, November 2006.

Luoma, S. N., (2008), "Silver Nanotechnologies and the Environment Old Problems or New Challenges?", PEN 15. Washington, DC: Project on Emerging Nanotechnologies, Woodrow Wilson International Center for Scholars.

Mamikunian, V., (2007), "Investor Enthusiasm for Nanotech Opportunities in Electronics, "Lux Research Inc. 3-15

Mousa S.A., & Linhardt, R., (2010), "Silver Nanoparticles as Anti-Microbial", US200100317617A1, 2010.

Mueller, N., & Nowack, B., (2008), "Exposure Modeling of Engineered Nanoparticles in the Environment", Environ. Sci. Technol., 42, pp. 4447–4453

Nadia, D. M., Catherine, B., Francis, C., BOUDHRIOUA Nourhène, B., Nabil, K., (2011) "Moisture Desorption Isotherms, Isosteric Heats of Desorption and Glass Transition of Fresh Pear and Apple: Experimental and Mathematical Investigation, European Drying Conference –Eurodrying 2011, pp. 1-4 Available At http://www.uibcongres.org/imgdb/archivo_dpo11056.pdf

NANO-CARE® Fabric Protection Named As One of TIME Magazine's Coolest Inventions of the Year, Greensboro, NC, November 18, 2002 Available At Http://Www.Bopuniforms.Com/Images/Nanocare.Pdf

Nischala, K., Rao, T. N., & Hebalkar, N., (2011), "Silica–Silver Core–Shell Particles for Antibacterial Textile Application", Colloids and Surfaces B: Biointerfaces, 82, pp. 203–208

Nowack B., Krug H.F., & Height M.J., (2011), "120 Years of Nanosilver History: Implications for Policy Makers", Environ. Sci. Technol. 45, pp. 1177 –1183.

Nowack, B., (2010), "Nanosilver Revisited Downstream", Science, 330, pp. 1054-1055.

Nowack, B., Krug, H. F., & Height, M., (2011), "120 Years of Nanosilver History: Implications for Policy Makers", Environ. Sci. Technol. 45, pp. 1177-1183.

Oh S.G., (2002), "Technology Using Sulfur Compounds for Increasing Antibacterial Property/Sterilizing Power of Silver Nano Particle", KR2002043499A, 2002.

Panyala, N. R., Pena-Mendez, E. M., & Havel., "Silver or Silver Nanoparticles: A Hazardous Threat to the Environment and Human Health?", (2008), J. Appl. Biomed., 6, pp. 117-129

Powell M.., (2011), "Silver: Miraculous Cure-All or Toxic Heavy Metal? A Historical Review of Silver's Harmful Effects on Humans," Available At: Http://Www.Nanoceo.Net/Files/Silver_Magic_Cure-All_Or_Toxic_Heavy_Metal.Pdf

Powers C. M., (2010), "Developmental Neurotoxicity of Silver And Silver Nanoparticles Modeled in Vitro and in Vivo", Ph.D. Dissertation, Department Of Pharmacology & Cancer Biology, Duke University, Durham, North Carolina.

Rantanen, K. L., & Domb, E., (2002), "Simplified TRIZ," St. Lucie Press, New York

Rebitzera, G., Ekvallb,T., Frischknechtc, R., Hunkelerd, D., Norrise, G., Rydbergf, T., Schmidtg, W.-P., Suhh, S., Weidemai, B.P., & Pennington, D.W., (2004), "Life Cycle Assessment Part 1: Framework, Goal and Scope Definition, Inventory Analysis, and Applications", Environment International 30, 701– 720 .

Ross, S., Evans, D., & Michael Webber, M., (2002), "How LCA Studies Deal with Uncertainty", Int J LCA., 7 (1), pp. 47 – 52

Samuel, R., Almedom, A.M., Hagos, G., Albin, S., Mutungi, A., (2005), "Promotion of Handwashing as a Measure of Quality of Care and Prevention of Hospital-Acquired Infections in Eritrea: The Keren Study", Afr Health Sci., 5(1), pp. 4–13. PMCID: PMC1831903

SchaFer B., Tentschert J., & Luch A,. (2011), "Nanosilver in Consumer Products and Human Health: More Information Required!", Environ. Sci. Technol. 45, pp. 7589-7590.

Schilthuizen, S., (2009), "Smart Textiles Enabled by Nanotechnology, RFID and Sensor Technology", SCINT, October 2009 Available At http://Www.Scint.Nl/Docs/Smarttextilesscint.Pdf

Schneider R, (2008), Textile International Enterprises, Presented At Marketing In The United States, December 5 , 2008 Available At http://Export.Textiles.Org.Tw/Doc/(1)Marketing%20in%20the%20US%202011.Pdf

Sciubba, E., & Göran Wall, G., (2007), "A brief Commented History of Exergy from the Beginnings to 2004", Int. J. of Thermodynamics, 10 (1), pp. 1-26, ISSN 1301-9724

Senge, P.M. & Carstedt, G., (2001), "Innovating Our Way to the Next Industrial Revolution," MIT Sloan Management Review, Winter, 2001, pp. 24-38.

Senjen R., & Illuminato I., (2007), "Nanosilver – A Threat To Soil, Water and Human Health?" Friends Of Earth. Available At: www.Foeeurope.Org/Activities/Nanotechnology/Documents/Foe_Nanosilver_ Report.Pdf

Senjen, R., & Illuminato, I., (2009), "Nano & Biocidal Silver – Extreme Germ Killers Present a Growing Threat to Public Health," Friends of Earth, June 2009.Available at: www.foeeurope.org/activities/nanotechnology/Documents/FoE_Nanosilver_ report.pdf

Serban, D., Man, E., Ionescu N., & Roche, T., (2004) "A TRIZ Approach to Design for Environment," (D. Talab and T. Roche eds.), Product Engineering, pp. 89–100.

Shin, N-M., (2008), "Obesity in South Korea", School Of Nursing, Korea University, Available At Http://U21health.Mty.Itesm.Mx/Sites/Default/Files/27_0.Pdf

Sifniades, S., Levy, A. B., & Hendrix, J. A. J., (1999) "Processes For Depolymerization Nylon-Containing Whole Carpet To Form Caprolactam", US Patent 5,929,234, July 27, 1999. US Patent 5,932,724, August 3, 1999.

Silebi, C. A., & Schiesser, W. E., (1992; 2011), "Dynamic Modeling of Transport Process Systems", (ISBN: 0126434204 / 0-12-643420-4/0126434204), Elsevier Ltd., NY.

Sotiriou G.A., & Pratsinis S.E., (2010), "Antibacterial Activity of Nanosilver Ions and Particles ", Environ. Sci. Technol., 44 (14), pp. 5649–5654.

Stensberg M.C., Wei Q.S., Mclamore E.S., Porterfield D.M., Wei A., & Sepúlveda M.S., (2011), "Toxicological Studies On Silver Nanoparticles: Challenges And Opportunities In Assessment, Monitoring And Imaging", Nanomedicine, 6(5), pp. 879-89.

Sun, C.Q., (2007) "Size dependence of nanostructures: Impact of bond order deficiency", Progress in Solid State Chemistry, 35, pp. 1-159

Tan W., Santra S., Zhang P., Tapec R., & Dobson J., (2003), Coated Nanoparticles, US6548264B1, 2003.

Terninko J., Zusman A., & Zlotin B., (1998), "Systematic Innovation – Introduction to TRIZ (Theory of Inventive Problem Solving)", Russia, CRC Press.

Tolaymat T.M., El Badawy A.M., Genaidy A., Scheckel K.G., Luxton T.P., & Suidan M., (2010), "An Evidence-Based Environmental Perspective Of Manufactured Silver Nanoparticle In Syntheses And Applications: A Systematic Review And Critical Appraisal Of Peer-Reviewed Scientific Papers", Sci. Total Environ., 408(5), pp. 999-1006.

U.S. Commercial Service, Asia-Pacific Cosmetics and Toiletries Market Overview March 11, 2011. Available at http://Export.Gov/Hongkong/Build/Groups/Public/@Eg_Hk/Documents/Web content/Eg_Hk_039451.Pdf

United States Environmental Protection Agency, www.epa.gov/lean, August 2009, EPA-100-K-09-006, http://www.epa.gov/lean/environment/toolkits/professional/resources/Enviro-Prof-Guide-Six-Sigma.pdf

Uznanski, P., & Bryszewska, E., (2010), "Synthesis of Silver Nanoparticles from Carboxylate Precursors under Hydrogen Pressure", J Mater Sci., 45, pp. 1547–1552

Vaidyanathan, R., Gopalram, S., Kalishwaralal, K., V. Deepak, V., Pandian, S. R. K., & Gurunathan, S., (2010), "Enhanced Silver Nanoparticle Synthesis by Optimization of Nitrate Reductase Activity," Colloids and Surfaces B: Biointerfaces 75, pp. 335–341

Volpe, R., (2010), "Letter to EPA Regarding the EPA Nanosilver Scientific Advisory Panel Report," Silver Nanotechnology Working Group (SNWG), February 2010

Wall, G., (1988), "Exergy Flows in Industrial Processes", Energy, 13(2), pp. 197-208.

Wijnhoven S.W.P., Peijnenburg W.J.G.M., Herberts C.A., Werner I. Hagens, W. I., Oomen, A. G., Heugens, E. H. W., Roszek, B., Bisschops, J., Gosens, I., Van De Meent, D., Dekkers, S., De Jong, W. H., Van Zijverden, M., Sips, A. J. A. M., & Robert E. Geertsma, R. E., (2009) "Nano-Silver – A Review Of Available Data And Knowledge Gaps In Human And Environmental Risk Assessment Nanotoxicology", 3(2), pp. 109-138.

Winslow, C.-E. A., (1920), "The Untilled Fields of Public Health," Science, n.s. 51, pp. 23

Wong Y.W.H., Yuen C. W. M., Leung M. Y. S., Ku S.K.A., & Lam, H.L.I., (2006), "Selected Applications of Nanotechnology in Textiles," AUTEX Research Journal, 6 (1), pp. 1-8 Available at http://Www.Freewebs.Com/Jayaram-Co/Doc/Selected_Appz_Of_Nanotechnology_In_Textiles.Pdf

Yang, J. H., Park, S. E., Chung, I. J., & Kwon, T. Y., (2010), 'The Silver City," Fall 2010 Senior Course (KU 3176) Project, Department of Materials Chemistry and Engineering, Konkuk University, Seoul, Korea.

Yang, W. D., (2006), "Nano Silver Container For Keeping Disinfectants Including Gauze, Alcohol, Hydrogen Peroxide And The Like", KR2006102451A, 2006.

Yu, I. J., (2008), "Subchronic Inhalation Toxicity Evaluation of Silver Nanoparticles," KEMTI, 5th Korea-US Nano Forum, April 17-19, 2008.

Zhang, J. J., Gurkanb, Z., & Jargensen, S. E., (2010), "Application of Eco-Exergy for Assessment of Ecosystem Health and, Development of Structurally Dynamic Models", Ecological Modelling 221, pp 693–702

Zhang, Q-Y., (2002), "Multiple Objectives Application Approach to Waste Minimization", Journal of Zhejiang University, Science. 3(4), pp. 405-411.

Zhao W., Li L H., & Danzeng LB., (2010), "Study Of Size Effect on the Conductivity of Nano-Silver Colloids," Microwave and Millimeter Wave Technology ICMMT Proceedings. Chengdu, May 8-11, 2010.

Zhu H., & Zhu L., (2004), "Anti-Coagulation Nano Silver Antibiotic Dressing", CN1473553A, 2004.

Zhu H., Zhu L., (2002), "Method For Preparing Micro Powder Containing Anti-Agglomerated Nanometer Silver, Micro Powder Produced By The Method And Its Application", WO2002090025A1, 2002.

Section 3

Health Systems

11

New Challenges in Public Health Practice: The Ethics of Industry Alliance with Health Promoting Charities

Nathan Grills
University of Melbourne
Australia

1. Introduction

In an increasingly market driven society, characterised by neoliberal economic policies and promotion of free trade, powerful multinationals have become significant actors, for good and bad, in global public health. These powerful multinational companies are using increasingly sophisticated marketing strategies not only to promote products - some of which are deleterious to health – but also to lobby against public health initiatives that threaten their profit. Should public health practitioners cooperate with, or even attempt to coopt, these powerful organisations in an endeavour to promote health? Although this seems to be an increasing trend one must remain cogniscent that these companies will promote profit at the expense of health and often they are more effective at coopting health causes for their profit driven purposes than health causes are at coopting them for public health ends (Wright 2010).

In particular, this chapter explores how sponsorship of charities by corporates is actually a form of advertising that, when unhealthy products are promoted, can damage public health. The favoured approach by industry to minimise negative impacts of such advertising is via self regulatory codes. However, in Australia and elsewhere, these have by and large failed (Handsley E, Nehmy C et al. 2007; Ofcom 2008; National Preventative Health Taskforce 2009, p151; World Advertising Research Centre 2009). For example, in Australia the voluntary self regulatory policy to limit advertising of unhealthy products to children, called the Quick Service Restaurant Industry (QSRI), has resulted in no meaningful change since being introduced in 2009. The New South Wales Cancer Council concluded that "Children's exposure to unhealthy fast-food advertising has not changed following the introduction of self-regulation" (Chapman, Hebden et al. 2011). Is it time for policy makers to impose limits on the promotion of unhealthy products in order to protect the health of the public.

It is not so surprising that self regulation initiatives fail as it is counter intuitive, and against shareholder interests, for a profit seeking industry to minimise profit through self regulation. For example, arround 50% of profit from gambling comes from those who are being harmed by the product: "problem gamblers" (The Public Health Association of Australia 2008). Therefore limiting advertising of unhealthy products in order to remove damage to health would threaten the viability of such industries and that outcome is definitely not in the interests of the shareholders!

In relation to corporate funding of health charities this seems to be entirely unchecked by either government or industry self regulation. At the very best this approach involves fundraising for a good cause that would otherwise be underfunded, and no doubt the charities themselves have no other motivation than to see their important cause supported. However, cynics of Corporate Social Responsibility (CSR) would argue that the ultimate goal for industry is profit, or at the very least trying to mitigate criticism of the organisation (Wright 2010). At its most sinister, might CSR involve an ethically questionable model whereby the charity is exploited to promote a company whose product is deleterious to health? This chapter describes how unethical behaviour increases along a spectrum when using charities to advertise by:

1. Funding a charitable cause in order to advertise a product
2. Funding a children's charity to advertise a product
3. Funding a children's charity to promote a product that causes harm
4. Funding a children's charity to promote a product that causes the very illness that the charity seeks to respond to
5. Funding a children's charity to promote a product that causes the very illness that the charity seeks to respond to, and use this sponsorship to attain the high moral ground and lobby against public health approaches to address the public health problem

Although there are various international case studies one could cite (see www.cmaj.ca/cgi/content/full/cmaj.110085/DCI for a list) this chapter unpacks four examples demonstrating potentially unhealthy alliances where industry has seemingly coopted children's charities and public causes in order to sell a product that damages health. This discussion attempts to raise awareness about such subtle marketing and intends to help readers discern what might be appropriate and inappropriate use of charitable causes.

Ultimately, we would hope that reading this chapter leads the reader towards taking action to protect our most vulnerable consumers from powerful industry interests. The chapter finishes by exploring how those in public health can creatively engage with this issue and respond by even using many of the same tactics utilised by companies whose products damage health.

2. What is an acceptable form of company sponsorship of health charities?

Advertising and marketing is very effective at selling 'goods', but these goods are not necessarily good. In the area of marketing Energy Dense and Nutrient Poor Food and Beverages (EDMPFB) various international reviews have concluded that heavy marketing is likely to have deleterious effects on children by encouraging products high in salt, sugar and fat (World Health Organization 2003; Livingstone 2006; National Preventative Health Taskforce 2009). Accordingly, in many countries, various codes exist to regulate marketing. However, Australian restrictions have been largely voluntary self-regulated codes which have failed to prevent ethically questionable advertising, such as advertising to children (Hebden, King et al. 2011) (Chapman, Hebden et al. 2011).

On the surface it seems acceptable, or even desirable, that a company whose product causes damage should contribute to alleviation of the same damage. Such is the basis for carbon credits and taxes whereby companies contributing to carbon production may choose, or be required, to contribute towards mitigation of the problem to which they contribute. An

example from the health field is DrinkWise, a charity funded by the alcohol industry, aiming to shape "a healthier and safer drinking culture in Australia where drinking to excess, or drinking too young, is considered undesirable". Such recompense is desirable if the aim and the effect is mitigation.

However, sponsorship is ethically tenuous when a company whose product potentially causes illhealth, assists victims, and in doing so advertises the very product that caused the illhealth. Libertarians would argue that companies should be entitled to pursue such strategies, as informed adults are capable of discernment and can decide accordingly.

However, there is a flaw in the assumption of consumers being fully informed or having the necessary agency to make such distinctions. For example, are people aware that DrinkWise is industry sponsored and have been accused of promoting the very products that caused the harm they are seeking to mitigate? The consumer may be unable to make a fully informed choice if the true identity of the organisation is unclear. If company X sponsors a charity Y which addresses the ill-health caused by the same company X, then consumers should surely be informed that charity Y is supported by Company X whose product causes the ill health.

However, it seems more ethically objectionable when a company whose product could harm children, then assists the children who might be harmed, to make these children consume more of their potentially harmful product. Not only is a potentially harmful product being advertised to children, but the immature target audience, unaware of the danger of the product being marketed, could be influenced to view it as harmless or even good. Acknowledging the effect of advertising unhealthy products to children, Australia's National Public Health Task Force (NPHTF) recommended phasing out of "premium offers, toys, competitions and the use of promotional characters, including celebrities and cartoon characters, to market EDNP food and drink to children across all media sources". Similar moves have been initiated in other countries (Handsley E, Nehmy C et al. 2007; Ofcom 2008; National preventative health taskforce 2009, p151; World Advertising Research Centre 2009) (World Health Organization 2003; Livingstone 2006): Sweden and Norway prohibit commercial advertising directed at children via television, and Quebec prohibits the use of any media (Handsley E, Nehmy C et al. 2007, p153).

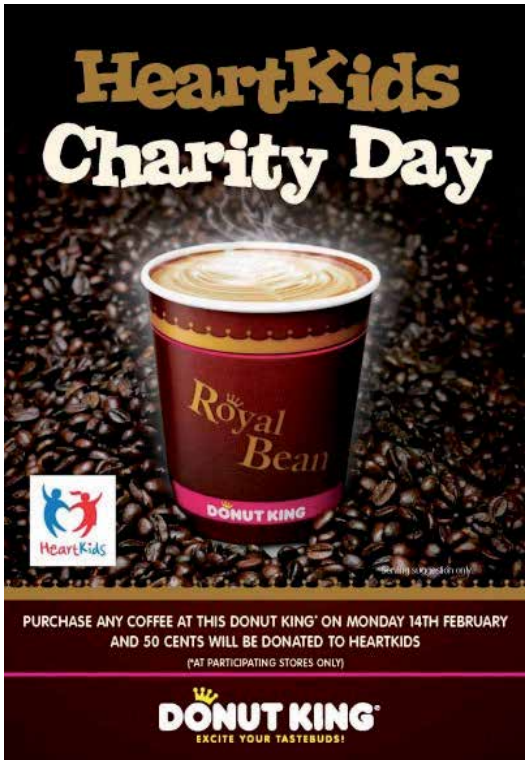
Perhaps even more insidious is where the funding of worthy charities creates an unhealthy alliance which allows a company to attain the moral high ground, and so limit their vulnerability to challenges regarding their unhealthy products and questionable practices. For example, if a policy-maker decided to limit a company's ability to inappropriately market their product to children, the company might then generate popular opinion against the politician with arguments like "this will undermine our ability to support children's charities such as the Ronald McDonald kids health truck!" (Prisk 2011). This supports a concept described in the literature where Corporate Social Responsibility, such as sponsoring a charity, is really about company credibility and positioning in order to benefit their bottom line (Wright 2010), or as Wright describes, limited to where it is profitable and often as a reaction to criticism of their product and practices. An example of ethically questionable practices might involve sponsoring a children's organisation to divert attention away from a product that potentially harms children.

I will outline four case studies which demonstrate that this practice might be more common than we perceive. Each case study may represent the intentional use of CSR to gain moral

high ground and sell potentially harmful products, or may be merely coincidental. Either way these practises need to be challenged.

2.1 Case study 1: The Donut King alliance HeartKids

The Donut King has become a regular supporter of HeartKids which is a foundation to support children with heart diseases and their families. On a single day in 2011 they kindly offered to give 50 cents of every purchase of a coffee from their fast food chain to the HeartKids charity. Of course this was promoted widely through adverts and in store promotion that cobranded Donut King products with the HeartKids logo (HeartKids 2011). Many readers would not initially discern any problem with Donut King supporting such a worthy charity, but the partnership warrants closer scrutiny.



Doughnuts are Energy Dense and Nutrient Poor Food and Beverage (EDNPFB) foods and such foods are linked with childhood obesity and cardiovascular disease later in life (National preventative health taskforce 2009). Whilst adults might be aware that doughnuts are potentially damaging EDNPFB and that sponsorship of a charity might actually be advertising, children may not be (HeartKids 2011). Additionally, children may be incapable of disentangling the apparent contradiction of an advertisement that links a fast food chain selling unhealthy food that can ultimately damage hearts, with a charity promoting healthy hearts in children! Is such advertising ethical if it exploits our most vulnerable community members: children and their health?

Indeed, such sponsorship can divert attention away from the potential harms of this company's EDNPFB whilst also attaining a moral high ground. That is, any challenge to the Donut King's charitable sponsorship - probably including this challenge - will immediately draw a response such as "Get a heart! Are you saying we shouldn't support HeartKids"! Interestingly, a medical colleague originally forwarded me this advertisement and encouraged us to visit Donut King to support HeartKids. I couldn't easily express my disapproval to her given that her child suffered congenital heart disease and would potentially benefit financially by the sponsorship.

The objection was not that children with congenital heart disease are harmed by Donut King's product and, indeed, most children with congenital heart disease need a high calorie intake. However, looking beyond the individual level, is there a population effect of normalising such unhealthy products? Through such sponsorship the Donut King is promoted as a good citizen who cares about health, and its products might be widely associated with a health cause, both of which may potentially increase sales of unhealthy products population wide. Secondly, even if the population effect is small, is it ethically appropriate to promote unhealthy products using vulnerable children to convey a message that this company cares about the very hearts that their product may damage?

2.2 Case study 2: McDonalds' alliances with the Royal Children's Hospital (RCH)

McDonald's relationship with the RCH Melbourne permits them to have a fast-food franchise on the hospital's grounds (Royal Children's Hospital 2010). The EDNPFBs that McDonalds promote are linked with childhood obesity and ill-health. In the famous UK libel case McDonalds sued two individuals for disseminating brochures claiming that McDonalds, amongst other things, was bad for health. The UK court of appeal found that "there is a respectable (not cranky) body of medical opinion which links a junk food diet with a risk of cancer and heart disease' and 'this link was accepted both in the literature published by McDonalds themselves and by one or more of McDonald's own experts and in medical publications of high repute'" (Judgement, p169).

Granted, McDonalds at RCH is much appreciated by parents and children alike, making the perfect sweetener for a child facing the trauma of visiting hospital. I also confess that as a father of a chronically unwell child, after leaving the ward at 10pm I have visited McDonalds to wind down in a friendly environment. However, I could have just as easily wound down in whatever cafe or restaurant was still open and accessible. Additionally, the government has stepped in to require the McDonalds at the new RCH to provide 80% healthy foods choices (green and amber) whilst restricting unhealthy food choices to 20% (Royal Children's Hospital 2010).

Yet the concern around McDonalds in the RCH is more complex than the negligible health impact on individual parents or children visiting McDonalds on a few random occasions whilst receiving care. There are ethical concerns about this alliance. Firstly, McDonalds can use their sponsorship to promote their brand name and unhealthy products to children and the wider community. This normalisation of EDNPFB consumption in the wider community is hazardous given that childhood obesity is approaching 30% in Australia. Secondly, is it ethically acceptable to allow our most vulnerable children to be exploited for the marketing of potentially unhealthy and harmful products? We allow these companies to promote an

image of a company which cares about the very children that their product may harm. Finally, is it acceptable to allow McDonalds or Donut King to attain the moral high ground by affiliating themselves with children's healthcare institutions and causes? It is very difficult to oppose unhealthy practices and products when these 'good corporate citizens' are seen to be promoting children's health.

So why do we allow an organisation whose product may damage children's health to sponsor our children's hospital? Are there other 'healthier' organisations which could support the RCH? I would like to reiterate Margaret Chan's challenge to such companies: "I would like to ask the food and beverage industries. Does it really serve your interests to produce, market, globally distribute, and aggressively advertise, especially to children, products that damage the health of your customers?"(Chan 2011). Again, such a case is difficult to sustain given outrage generated by threatening funds for children's healthcare (Prisk 2011). McDonalds are very safely on the moral high ground.

How can one argue against "supporting families" and "helping seriously ill children"?

2.3 Case study 3: Tattersall's alliance with the RCH

Would you allow the following company to speak to your kids when the company has a majority stake in an industry which:

- Makes more than 50% of their profit by trapping powerless addicts (Hancock, Schellinck et al. 2008)
- Profits from the most vulnerable and poorest
- Increases crime rates in the area (study reported in the Age, July, 2010)
- Impoverishes and breaks up thousands of families in Australia each year
- Is associated with one in five suicide attempts in patients presenting to the Alfred Hospital in Melbourne

Most responsible parents would not allow such companies to promote their product to their children, so why does RCH allow Tattersall's to do so in the Children's Hospital? A RCH Foundation report provides an answer:

> "The ongoing contribution of Tattersall's, one of the hospital's longest standing corporate partners, has reached a total of $8 million. Each and every dollar that comes into the Foundation represents both personal sacrifice and the affectionate regard that Victorians have for the Royal Children's Hospital." (Royal Children's Hospital 2010).

This outlines both the reasons for the link ($8 million) and also demonstrates the moral high ground obtained in that every dollar from Tattersall's "represents both a personal sacrifice and affectionate regard". How can one question a donor who has an affectionate regard for an important institution like the RCH? Such organisations can use their moral high ground to influence policy as was shown in the recent senate committee investigation into gaming where Woolworths, the biggest owner of electronic gaming machines, threatened that they would have to decrease their investment in the community if profits from electronic gaming machines were limited by legislation (Needham 2011 (Feb 11)). Perhaps the new RCH, "completed in 2011", might reconsider this unhealthy association?

The community needs to recognise that the products that Tattersall's promotes actually cause significant harm to the very society that it claims to be helping through its support of charities. One expert researcher in the field, Professor Charles Livingstone argues:

> "The problems of pokie gambling are not trivial. They include financial distress and ruin, bankruptcy, fraud, embezzlement, and theft and misappropriation of the funds, property and income of family, friends, employers and others. Gambling problems are also strongly associated with crime generally, family breakdown, divorce, the neglect and abuse of children, mental and physical illness, depression and anxiety, and not infrequently include suicide. The children of regular and problem gamblers are themselves significantly more likely to have a gambling problem than those of non-gamblers, and poker machine venues are most strongly concentrated in poorer suburbs"(The Public Health Association of Australia 2008).

However, despite Tattersall's association with such damage, it has represented itself by sponsoring the very society that it harms. Tattersall's do not only sponsor children's hospitals and hospital emergency departments in Australia but various sporting clubs in which our children participate. Many sporting clubs have become dependent on the revenue from sponsorship of Gambling agencies or from revenue from owning gaming machines. There is little doubt that allowing the gambling industry to operate in sports clubs exposes children to advertising and normalisation of such products. Do we need better protection? However, once again the gambling agencies have attained the moral high ground where clubs and supporters may well contest that the club depends on that revenue. If we ban the "Tattersals' sponsorship" then we risk accusations of compromising institutions that actually promote health. In Australia McDonalds has similarly inserted itself into the health DNA of our schools and youth clubs through sponsoring Auskick, kids sporting events and, at a higher level, sponsoring Australian international sporting teams such as the Australian Olympic Team. The Australian Olympic Team website allows McDonalds to boast "In Australia, we are very proud to be helping kids be active by supporting Little Athletics, Soccer, and Basketball in various states" (http://corporate.olympics.com.au/sponsor/mcdonalds)

2.4 Case study 4: Alcohol industry alliance with children's fundraising

There is good evidence that exposure to alcohol advertising shapes young adolescents' attitudes toward alcohol, their intentions to drink, and underage drinking behaviour (Martin 2002). Additionally, studies show that alcohol advertisements are often shown during the shows that target teens such as sporting events (Martin 2002). In Australia, thankfully, alcohol advertisements are no longer shown during children's programs. However, alcohol advertisements do still target youngsters during shows watched by large numbers of children such as sporting events. In 2002, in the US, over a billion dollars was spent to advertise alcohol on TV and around 22% of these advertisements were seen more by youth than adults (Center on Alcohol Marketing and Youth 2004).

Similarly to Tattersall's and McDonalds, alcohol companies do not only advertise through traditional media platforms. The alcohol industry has been a regular sponsor of sporting and charity events held through our schools. This can achieve a similar end to more traditional forms of advertising. A recent report by the Australian National Council on Drugs found that alcohol was often the focus of various fundraisers which include supporting wine "drives" conducted via newsletters, liquor "tasting events" on school premises, and alcoholic bottles featuring school logos. Dr Herron from the Australian National Council on Drugs states:

> "I think we all know subliminal messages have a huge impact on young people. Through attaching (fundraisers) to a school newsletter, we're legitimising them and saying it's all right for students to be transporting information about alcohol between the home and school." (Barry 2011)

Again it must be questioned if such charitable sponsorship is benevolent or little more than blatant advertising to adults and children by profit driven alcohol companies.

However, the involvement of charitable causes makes rational debates on this issue difficult to have. Few critics of advertising to children would doubt that well meaning parents and friends have the best intentions in raising funds for worthy causes. Indeed, the alcohol companies take advantage of this very fact to, once again, attain the moral high ground and an immunity to being challenged. Parents and friends become a powerful ally for the alcohol companies and might well defend the company by passing off the 'sponsorship' as harmless and merely for philanthropic purposes.

Similarly to the first three case studies, we again question if it is ethical to allow an alcohol company to link itself to a school when damage from youth alcohol usage is so prevalent and damaging throughout Australian society (Chikritzhs, Pascal et al. 2004). The list of damage caused to youth by alcohol is long and well established but can be best summarised by the fact that Alcohol accounts for 13 % of all deaths among people 14-17 years of age and in Australia, each week, one teenager dies and around 60 are hospitalized from alcohol-related causes (Jones, Chikritzhs et al. 2004; Clark, Thatcher et al. 2008). Teenagers without the benefit of good judgement from experience, are particularly vulnerable to alcohol related harm in a way that older drinkers may not be (Australian Medical Association 2009). Among young Australians, the most common causes of death and injury due to risky or high-risk drinking are road injury, suicide, and violent assault (Chikritzhs, Pascal et al. 2004). The Australian School Students' Alcohol and Drug Survey (hereinafter ASSAD)

highlights the extent of the alcohol problem amongst youth with 13% of children aged 16 year olds having drunk at dangerous levels in the past week (Centre for Behavioural Research in Cancer 2008).

As these statistics suggest, allowing companies to insert themselves into schools is self evidently unacceptable but the alcohol industry inserts itself more insidiously into children's health causes outside the school environment. The AMA documents how the increasingly sophisticated marketing of alcohol is aimed at attracting, influencing, and recruiting new generations of potential drinkers (Australian Medical Association 2009). One example is the targeting and supporting of not-for-profit Australian sporting clubs by alcohol companies. Children involved in these clubs grow up viewing alcohol advertisements and conceivably accepting the industry as an important part of their society and a promoter of good health. Nothing could be further from the truth given that alcohol is one of the greatest dangers faced by the young people of Australia! Furthermore, alcohol, from whatever perspective you look at it, is damaging to sporting performance so it is ironic, or maybe intentional, to link a health damaging product to health promoting activities. Instead such support of charities by the alcohol industry would seem to be another example of an industry injurious to children's health allying itself with children's charities in order to promote its product and image.

3. Why is it unethical?

To help determine if a sponsorship is ethical it is also worth referring to the stewardship model outlined in the report by the Nuffield Council on Ethics (2007). They concluded that in regards to the role of industry, the media and other parties, "businesses have obligations towards society. Many businesses already have social responsibility policies. Where industries fail to meet reasonable standards it is acceptable for the state to intervene through regulations (Paragraphs 2.47-2.50 and 3.41)". If the above case studies represent failure to meet acceptable standards, then should the state government intervene to limit this sponsorship?

Secondly, when promotion of unhealthy products involves children, through the use of children's health institutions and charities, the mandate for action is clearer. The stewardship model outlines "protecting and promoting the health of children and other vulnerable groups" as a high order principle that can justify limiting freedoms (Nuffield Council on Bioethics 2007). It would seem that allowing companies to exploit children's charities to promote harmful products would go against the stewardship model. In effect children represent a market failure due to imperfect information and information asymmetry as they are incapable of being fully informed consumers. We therefore have an ethical mandate to steward our most vulnerable by protecting them against exploitation.

Permitting companies to exploit children's charities and children's health services to promote harmful products might qualify as "behaviour harming others". According to J.S. Mill, in his famous volume 'On liberty', intervention by the state is only justified when behaviour harms others, as such sponsorship might do if it causes more consumption of the harmful products (Mills 1909). Similarly a recent article in the Lancet argues "Liberty should be restricted, in a liberal society, only when there is a clear and direct threat of harm to

innocent parties who cannot respond for themselves" (Finn and Savulescu 2011). There seems adequate evidence now that advertising of unhealthy products to children does cause harm to children who are incapable of responding.

The intervention ladder developed under the Nuffield Bioethics report holds that more intrusive interventions require stronger justification. Although the ethics of advertising to children is still being contested, we conclude that allowing companies promoting unhealthy products to link their product to health institutions or causes, is a justification for action.

Regulating advertising to children is gathering widespread support in Australia where consumers (or more accurately the parents of consumers) are tired of having to fight against blanket advertising to maintain healthy diets for their children. Key findings from a recent phone survey is South Australia were:

- 85% of consumers believe children should be protected from unhealthy food advertising.
- 93% of people were in favour of the government introducing stronger restrictions to reduce the amount of unhealthy food and drink advertising seen by children, with 79% strongly in favour.
- 86% of grocery buyers are in favour of a ban on advertising of unhealthy foods at times when children watch TV, with 70% strongly in favour.
- When asked what most commonly negatively impacted their children's food purchase requests, grocery buyers reported television commercials (36%) or toys and giveaways (24%).

(Cancer Council SA 2011)

Along these lines of protecting minors, the Gambling Regulation Act 2003 (Vic) would seemingly be justified in seeking "to ensure that minors are neither encouraged to gamble nor allowed to do so" (section 1.1.iib) (Victorian Government 2003). This act therefore challenges Tattersall's promotion of their brand at the RCH.

4. What action should be taken?

If such behaviour is proceeding with little regulation in many countries then what can be done? I suggest a similar approach to what has worked in previous campaigns such as the one to limit tobacco companies' right to advertise their harmful product. It has taken a concerted, multipronged and sustained campaign to undermine the supposed right to advertise this dangerous product. This included advertising, mobilising physicians around the cause, raising public awareness, undertaking research and advocating to the policy makers and key stakeholders.

Firstly we believe that public health practitioners and doctors should raise awareness of potentially unethical approaches. Doctors and the health profession more generally are still widely respected by the community. As professionals concerned for the health of those in our community we must be making efforts to protect the health of our most vulnerable. At the very least, awareness can be raised in the public health arena by writing to media outlets, journals and other fora in order to expose, or at least question, apparent unhealthy alliances. Such lobbying has been shown to be an important part of a

concerted campaign, and effective when it is part of a multipronged approach. The use of the new media is also important. A website called unhealthyalliances is under development. A campaign in Canada drew on a Facebook group to undermine the Burger King's placement of its product in a children's hospital.

In particular, health professionals should advocate for the banning of advertising of damaging products in children's hospitals and institutions where they work. After all, the problem is not primarily related to the companies, which are by nature profit driven. Instead the onus falls largely on the health organizations themselves where many of us work. We should be continually challenging our employers towards more ethical behaviour by dissuading them from accepting money from, and partnering with, companies whose products damage children's health. We should not accept ethical standards being compromised merely in order to finance health programs, buildings and services. In Toronto staff contributed to preventing the Burger King from continuing to operate at the Sick Kids Hospital. The group drew comments from physicians and health professionals to add pressure not to renew the Burger King's lease despite the $2.5million the Burger King had raised for the hospital (Farquharson 2011, March 20).

Physicians who sit on boards and advise on hospital governance issues need to avoid being complicit by not taking action. They can advocate for regulations and clauses to limit the food industry exploiting children. An article in the CMJ advises that at the very least "partnerships should comprise unconditional arm's-length grants with clauses limiting how corporations use health organization brands" (Freedhoff Y and PC. 2011). They warn that if we do not act we risk compromising health promotion goals by helping to promote unhealthy brands.

Awareness could be raised through counter advertising campaigns aimed at unravelling the unhealthy alliance between health charities and a company which promotes unhealthy products.

https://www.getup.org.au/campaigns/pokies-reform/grandfinal-ad/get-this-ad-on-the-air

The Get Up advocacy group has produced various counter advertising campaigns such as the one challenging the positioning of pokies in sporting clubs frequented by children. Such campaigns can be particularly effective but are often prohibitively expensive and risk defamation cases being brought against the group.

Given the significant power of the companies and their ability to scare journals, publishers and media formats, is it reasonable to revert to the type of tactics used in the Billboard Utilising Graffitists Against Unhealthy Promotions (BUGAUP) campaign? BUGAUP successfully countered tobacco advertising by adding counter slogans on the advertising by tobacco companies.

A similar idea was utilised in a recent campaign to expose Tattersall's unhealthy alliance with RCH. Members of the public used the Tattersall's advertising sign to educate the public about the hazards of gambling. Over a three month period eight messages were written, before the message was successfully conveyed and the RCH finally removed the sign. Whilst not advocating illegal graffiti, other legal forms of public health advocacy and protest on such important issues are important.

Graffiti on the sign week 2

No sign (week 9)

5. The bigger picture

This chapter has focussed on relevant examples from Australia where youth alcohol, childhood obesity and social problems from gambling are some of our most significant public health problems. However, this is a global problem and there are numerous examples from different countries where companies, whose product is harmful, link themselves to health organisations and health causes in order to mitigate their poor image or even to leverage support policies from these health organisations. The Canadian Medical Journal published a list of health organizations whose messages and reputations have been tarnished by partnerships with food companies (available at www.cmaj.ca/cgi/content/full/cmaj.110085/DCI).

The common message from all these examples is that we need to be cogniscent of this tendency whereby charities are utilised, or subverted, to ultimately sell unhealthy products to our children and community. We need to question if corporate sponsorship of charities is altruistic philanthropy or merely exploitation of charities to sell what can be dangerous products? Companies may not always act so insidiously but it should be remembered that they are ultimately accountable not to public health but to their shareholders who are concerned about the bottom line.

Beyond just attaining the moral high ground there is very real danger that such companies can use their support to pressure health institutions and policy making bodies to avoid implementing healthy policies that might damage the image, and profit, of the sponsoring company. Such unhealthy alliances also help the company to lobby against important health initiatives. An editorial in the CMJ describes how the CEO of Coca-Cola, Sandy Douglas, leveraged the company's relationship with the American Academy of Family Physicians to help make the case that soda taxes were unnecessary (Freedhoff Y and PC. 2011).

More recently there has been concern about corporate lobby power being brought to bear on multilateral UN agencies. An example of an unhealthy alliance with a multilateral is where UNICEF Canada, which amongst other things undertakes nutritional programs in developing countries, allowed its name to be used to promote Cadbury chocolate bars (Lancet 2010). Such partnerships are of growing concern given that changes to WHO funding mechanisms could see it receive more funding from, and work more closely with, the private sector. One commentator on the recent WHO reforms being discussed stated: "fears about WHO's independence remain as a result of the repeated calls for an increase in the role of the private sector and the possibility of funding from them". In effect the WHO would be opening itself up to a conflict of interest where the world's largest independent health watchdog and peak advisory and normative body in health, could receive funds from vested interests. If the food and beverage industry is allowed to become involved in sponsoring the WHO would it compromise the WHO's power to promote normative guidelines on obesity prevention which may involve setting limits on advertising to kids, and advising limits on salt/sugar/fat in certain foods? There is already such a precedent where food and beverage industry applied lobby pressure on powerful member states to oppose an evidence based guidelines around limits on sugar consumption.

6. Conclusion

We believe that there is an ominous, and largely unquestioned, trend for unhealthy products to be co-advertised with children's health services and charities. Whilst not accusing companies of inappropriate behaviour, this viewpoint challenges regulators and health institutions themselves to reconsider unhealthy alliances. These alliances potentially advertise unhealthy products to children, give companies that produce harmful products a moral high ground of supporting children's health, and ultimately undermine important health promotion messages. We argue that such activities are ethically questionable, and using a public health framework for ethics, warrant more intrusive regulations on advertising through our children's health institutions and charities.

In the new era of public health this issue must be dealt with effectively if we are to maintain our health levels and challenge the increasing double burden of infectious and non infectious diseases in the developing world. This is the new frontline in public health and we are currently lagging behind in this conflict. This chapter, it is hoped, has helped expose potential opposition to public health and this might serve as a call to action for public health practitioners and advocates.

7. Acknowledgments

Dr Bruce Bolam for helping develop the concept and reviewing a number of iterations of this paper along the way.

Prof Rob Moodie for modelling a public health advocate and for the encouragement to write about such issues.

8. References

Australian Medical Association (2009). Alcohol Use and Harms in Australia (2009) http://ama.com.au/node/4762.

Barry, E. (2011). "No place for booze in schools fundraisers, says Australian National Council on Drugs " The Herald Sun March 2.

Cancer Council SA (2011). Public supports tougher regulation of unhealthy food advertising. Adelaide.

Center on Alcohol Marketing and Youth (2004). Youth Exposure to Alcohol Ads On TV 2002.

Centre for Behavioural Research in Cancer (2008). Australian School Students' Alcohol and Drug Survey (ASSAD) Cancer Council Victoria.

Chan, M. (2011). Tackling food-related diseases: voluntary measures or regulation - carrot or stick? . The World Health Organization's global forum: Addressing the challenge of noncommunicable diseases, Moscow.

Chapman, K., L. Hebden, et al. (2011). "Advertising of fast food to children on Australian television: the impact of industry self-regulation." Medical Journal of Australia 195(1): 20-24.

Chikritzhs, T., P. Pascal, et al. (2004). "Under-Aged Drinking Among 14-17 Year Olds and Related Harms in Australia, National Alcohol Indicators." National Drug Research Institute, Curtin University of Technology Bulletin No.7.

Clark, D., D. Thatcher, et al. (2008). "Alcohol, psychological dysregulation and adolescent brain development." Alcoholism Clinical and Experimental Research 32(3): 375-385.

Farquharson, V. (2011, March 20). Burger King loses foothold at Sick Kids. The Globe and Mail. Toronto.

Finn, A. and J. Savulescu (2011). "Is immunisation child protection?" Lancet 378(9790): 465 - 468.

Freedhoff Y and H. PC. (2011). "Parnterships between health organisaitons and the food industry risk derailing public health nutrition (editorial) " CMAJ 183(3).

Hancock, L., T. Schellinck, et al. (2008). "Gambling and corporate social responsibility (CSR): Re-defining industry and state roles on duty of care, host responsibility and risk management." Policy and society 27: 55–68.

Handsley E, Nehmy C, et al. (2007). "Media, public health and law: A lawyer's primer on the food advertising debate. ." Media and Arts Law Review 12(1): 16.

HeartKids. (2011). "Donut King supporting heartkids on valentines day." from http://www.heartkidsvic.org.au/index.php/state/news_item/donut_king_supporting_heartkids_on_valentines_day/

Hebden, L., L. King, et al. (2011). "Advertising of fast food to children on Australian television: the impact of industry self-regulation." Med J Aust 195(1): 20-24.

Jones, P., T. Chikritzhs, et al. (2004). "Under-Aged Drinking Among 14-17 Year Olds and Related Harms in Australia, National Alcohol Indicators." National Drug Research Institute, Curtin University of Technology, Perth. Bulletin No.7.

Lancet (2010). "Trick or treat or UNICEF Canada." Lancet 376: 1514.

Livingstone (2006). New research on advertising foods to children - an updated view of the literature, in television advertising of food and drink products to children. London, Office of Communications.

Martin, S. (2002). "Alcohol Advertising and Youth." Alcoholism: Clinical and Experimental Research 26( ): 900-906.

Mills, J. (1909). On liberty P. F. Collier & Son.

National preventative health taskforce (2009). Australia: the healthiest country by 2020 National Preventative Health Strategy – the roadmap for action. Canberra, Commonwealth of Australia.

Needham, K. (2011 (Feb 11)). Pokies 'just like burgers'. The Age. Melbourne.

Nuffield Council on Bioethics (2007). Public health: ethical issues. London, Nuffield Council on Bioethics.

Ofcom (2008). Changes in the nature and balance of television food advertising to children: A review of HFSS advertising restrictions. London, Office of Communications.

Prisk, T. (2011). Truck drives access to health care. Centralwesterndaily.

Royal Children's Hospital. (2010). from http://www.newrch.vic.gov.au/Shopsservicesandamenities.

The Public Health Association of Australia (2008). Gambling and Health policy. Australia.

Victorian Government (2003). Gambling Regulation Act Australia

World Advertising Research Centre. (2009). "US government to scrutinise food marketing to children." Retrieved 1 May 2010, from www.warc.com/news/topnews.asp?ID=24840. .

World Health Organization (2003). Diet, nutrition and the prevention of chronic diseases. Report of a joint WHO/FAO expert consultation. W. T. Series. Geneva, World Health Organization. 916.

Wright, K. (2010). "Corporate Social Responsibility: A Review of the Literature." The higher education academy 19(24).

12

Primary and Hospital Healthcare in Poland – Organization, Availability and Space

Paweł Kretowicz and Tomasz Chaberko
Jagiellonian University
Poland

1. Introduction

Spatial distribution and location of healthcare facilities have been long acknowledged as main interests of Polish medical geography, although most research done dates back to the late 1980s and early 1990s (Mazurkiewicz, 1994; Michalski, 1999). These include e.g. some renowned studies of health services in Warsaw (Grochowski, 1988; Malczewski, 1989). Unfortunately, healthcare accessibility and availability have not been widely explored by geographers in the 2000s; thus marginalized in spatial sciences, has been detained by other disciplines such as public health (see Chawla et. al, 2004).

In 1952 geography of health was officially recognized and incorporated into geographical sciences by Medical Geography Committee operating within the structures of International Geographical Union. At the time, geography of health endeavored to investigate geographical factors of causes and consequences related to changes in population health status and morbidity. Presently, this subdiscipline consists of two distinctive strands: the spatial distribution of disease and death, and the geographical complexities surrounding the provision, access to and inequality of health care (Kearns & Gesler, 2002; Parr, 2003). Hence, most researchers clearly distinguish geography of healthcare as a domain that focuses on spatial accessibility of health services through the lenses of their distribution, demand, supply, utilization and planning (Mayer, 1982; Moon et. al., 1998). Moreover, geography of health care has evolved to investigate how medical resources meet population needs in space (Rosenberg, 1998; Kearns & Moon, 2002). Irrespective of such collaborative approaches some spatially-aware researchers frequently explore spatial and non-spatial factors underlying healthcare accessibility (Haynes, 2003; Wang & Luo 2005; Unal et. al, 2007).

The most important regulation of Polish healthcare system guarantees equal access for everyone, which is directly declared in the Polish Constitution, Article 68, Paragraph 2: „*Equal access to health care services, financed from public funds, shall be ensured by public authorities to citizens, irrespective of their material situation. The conditions for, and scope of, the provision of services shall be established by statute*". As suggested here, equal access refers to free utilization of health services. Although provided and financed by the state, health services should be congruent with other dimensions of accessibility. These are: affordability, accommodation, acceptability, availability and spatial accessibility (Penchansky & Thomas, 1981). First three dimensions can be viewed as non-spatial, however planning and fund distribution on healthcare in particular regions and counties should be based on potential

and actual/expected population needs (Guagliardo, 2004). The notions of availability and accessibility describe the relationship between location of healthcare facilities and patient residence. Availability reflects an assessment of how the volume and type of existing services (and resources) reflect the clients (patients) volume and types of needs (Joseph & Phillips, 1984). Spatial accessibility refers to distance, travel time, cost and modes of transportation.

The most important barrier to egalitarian conditions of healthcare utilization includes much higher demand of medical services as compared to the supply. This demand increases along with economic growth and development of new technologies. The existing medical resources (especially in secondary – specialized care and tertiary – hospital care) can no longer meet the needs of all patients simultaneously, nor do the financial resources can be distributed across all in need. As a consequence, long lines to the specialist offices discourage the sick and make them shift from public to non-public healthcare facilities. Worse still, heath services offered in non-public facilities are not always refunded by the National Health Fund (Narodowy Fundusz Zdrowia – NFZ – the institution responsible for redistribution of insurance contributions); thus patients have to pay for services. Furthermore, annual budgetary limits to health services also constrain access to healthcare and contribute to long wait times. These restrictions result from new technologies in medicine and pharmacy, which provide more efficient, but expensive medical equipment, treatments and medicines. In 2009 only 5% of patients generated no fewer than 60% of the total expenditures on services guaranteed by the NFZ (Ruszkowski, 2010).

Commercialization in Polish healthcare system has progressed dynamically since the early 1990s. Presently, about 75% of general practitioner offices and 82% of specialist offices in Poland operate as non-public facilities. In case of hospitals this proportion in 2010 amounts to 35% with a total number of hospital beds in non-public facilities reaching 32.8 thousands (16% of all hospital beds in Poland). This indicates that the majority of commercialized hospitals comprise relatively small facilities whereas the largest ones remain either state or province-owned.

2. Data sources and setting

According to Kaczmarek (2007) the availability of medical services depends on the volume and structure of current resources (e.g. number of medical doctors, nurses, hospital beds, medical equipment). Unfortunately, under the socialist rule before 1989 access to data concerning material and personal resources in healthcare was very limited. Neither existing registers nor Central Statistical Office was capturing data for all of the medical specialties. Besides, official figures were often aggregated and presented as simple classifications (Dziubińska-Michalewicz, 1994). Presently, Central Statistical Office (Local Databank, www.stat.gov.pl) provides scarce information concerning healthcare facilities (total number), although most is available on a municipal level. This data is divided by the ownership (public and non-public), type of care (primary and hospital) and utilization (crude number of consultations with physicians). County-level data includes the number of public and non-public hospital beds and the number of hospitalized patients (until 2003). In order to assess the distribution, organization, ownership and medical staff in healthcare facilities across the country the best databases offer the Register of Health Care Units (Rejestr Zakładów Opieki Zdrowotnej, www.rejestrzoz.gov.pl) and Central Register of

Health Professionals (Centralny Rejestr Lekarzy, www.nil.org.pl). These sources are extremely useful in geographical analyses of such subjects as:

a. location and organization of healthcare facilities
b. location of new facilities, changes of ownership (since 2004)
c. spatial accessibility and availability of health professionals by specialty
d. number and structure of hospital beds with respect to potential health needs

Center of Health Information Systems operating within the Ministry of Health is the main body responsible for data capture and storage. Annual Survey Programs of Official Statistics include public health data, which obligates every healthcare institution to send the required information to the Center monthly or annually. The records collected facilitate accurate analyses of medical resources and their utilization (data available on a municipal level). For example, MZ-88 and MZ-89 forms are used to collect data about medical staff employed in each health care unit in the whole country. Similarly, forms MZ-11 to MZ-15 include the data about number of patients and consultations with physicians by date, patients' age and sex, type of ailment etc. Analogous information about hospital care can be derived from MZ-11Szp forms provided by every facility in the country (by both facility location and patient residence). Unfortunately, not all hospitals follow this regulation what would result in strong underestimations (no data for one hospital in an area) if any geographical analysis was conducted. All of the data sheets are also sent to Centers of Public Health located in 16 provincial seats. Until recently, these institutions operated as separate entities, but they have now been incorporated into Health Departments of Province Offices. Administrative division of Poland authors wish to refer to in this study is presented below (Figure 1). Poland is divided into 16 provinces (49 until 1999), 314 land and 65 urban counties, 2478 self-governed municipalities.

Source: authors' own work

Fig. 1. Administrative division of Poland since 1999.

This study employs the data extracted from both Register of Health Care Units and Central Statistical Office. According to *The Act of 30th September 1991 on Health Care Units* (*Ustawa z dnia 30 września 1991 o zakladach opieki zdrowotnej*; Dz.U. 1991, Nr 220, Poz. 1600.) register entry is tantamount to official permission to run a medical office or health center (fines are levied upon those unregistered). Register of Health Care Units includes detailed information about facility address, location of its branches, legal foundations, organizational structure, type of medical specialty, number of beds in each ward (for hospitals and other inpatient clinics). Unfortunately, some information found in the Register turns out to be unreliable as not all specialist offices are included into the computer database (regardless of declared trustworthiness by the Center of Health Information Systems). For this reason, authors decide to take into consideration only primary and hospital healthcare. In spite of clear attempts to enhance data availability, a lack of necessary information provided by Polish statistical institutions is considered as a major limitation for health-related research and medical geography in particular. The main indicator of primary healthcare availability utilized this study includes practitioner's office to population ratio. In case of hospital care this measure comprises the number of hospital beds to population ratio.

3. Organization of healthcare system in Poland after World War II

Contemporary spatial organization of healthcare system in Poland has been shaped by historical determinants, healthcare model employed by the politicians as well as recent socio-economic processes. Under socialist rule, as in many other sectors of the national economy, management and planning in healthcare fell under central authorities. *The Act of 28th October 1948 on Collective Health Care Centers and Planned Economy in Healthcare* (Ustawa z dnia 28 października 1948 r. *o zakładach społecznych służby zdrowia i planowej gospodarce w służbie zdrowia*, Dz.U. 1948 Nr 55, Poz. 434.) virtually barred local governments and territorial health care centers from making consecutive decisions in healthcare organization and planning. Every resolution must have been first discussed and then approved by the district departments of national administration. Former healthcare system was organized in conjunction with the administrative division of the country. The provincial hospitals (so-called integrated provincial hospitals – Wojewódzki Szpital Zespolony) were most privileged as they offered the widest variety and highest quality of medical services. As a result, inequalities in spatial distribution of tertiary care increased significantly and favored these provinces with the largest district/regional facilities. Moreover, this gap widened after locations of some institutions in accordance with the political and military will of the Warsaw Pact (Ruszkowski 2008). Such locations were justified ideologically as communistic authorities were determined to arrange sufficient hospital infrastructure for the army in case of war (anticipated World War III). Hence, oversized and strategically-located institutions were being constructed across the entire country, but chiefly in the west of Poland. Consequently, large hospital facilities usually exceeded the needs of local population. A good example of such location could be Stanislaw Staszic's county hospital in Piła (Wielkopolskie Province), of which construction began in 1977 (Photo. 1.)

The number of beds in Stanislaw Staszic's Hospital in Piła peaked in 1992 when comprised as many as 726 beds. This number greatly exceeded local demand and was gradually reduced down to 601 beds in 2010; thus maladjustment of hospital size to the population needs was evident. The same problem concerned the spatial distribution of expenditures on healthcare financed both by the national (institutions of a nationwide range) and the

provincial budgets (institutions of a regional range). These expenditures were allocated without any regard to spatial distribution of population demand and healthcare utilization. All in all, fund distribution was organized by extrapolation of expenditures incurred in the previous year, providing their maximization with no rights to transfer any expenses for the next year (Curtis & Malczewski, 1990). Such management was ineffective, inadequate to the current social expectations and put numerous health care units in financial hardship. After the political transformation of 1989, this extensive policy led to shortages of personnel, medical equipment, medicines and favored corruption (Millard 1995).

Source: googlemaps.com

Photo 1. Stanislaw Staszic's county hospital in Piła.

The aforementioned system of healthcare managed and financed by the state budget and based on the lack of private health care institutions is called the Semashko model. This model, criticized for the extensive allocation of funds, dominated Poland and other socialist countries in the second half of the 20th century. Healthcare institutions were utilized only by patients who resided in the preventive-therapeutic districts (embracing from 30 000 to 150 000 inhabitants). Nonetheless, the basic units were called micro-district and embraced from 3 000 to 6 000 inhabitants. These units had to possess at least one physician in service and cover at least one village or district (borough) (Kaser, 1976). Moreover, health services for certain social groups were organized by completely different public bodies e.g. enterprise health services, railroad health services, Ministry of the Interior Affairs and Ministry of Defense health services (Grochowski, 1988). In 1972, the integrated health care management units - ZOZ (Zespół Opieki Zdrowotnej) were established. These entities were responsible for management of hospitals, outpatient clinics, specialist and primary health care as well as some social services. In 1991 health care units replaced integrated health care managements units, but retained the same acronym (ZOZ – Zakład Opieki Zdrowotnej). The Semashko model operated until 1999, when, in the aftermath of reforms in the Polish healthcare, insurance contributions were introduced. By this means, a transition from budgetary to insurance model put healthcare system in Poland on different tracks leading to the Bismarck model (social insurance model). This model introduces mandatory insurances, free choice of service and insurance provider as well as contract-based organization of healthcare system. New model alters spatial patterns of

healthcare utilization by the patients, who are now allowed to choose health professional and health care institution wherever they wish including locations outside their area of residence. Irrespectively, the system present in Poland is now criticized as the individual contributions remain involuntary, do not depend on individual decisions and the insured have no influence on the quality of service received (Siwińska et. al., 2008).

Presently, changes in Polish healthcare system to some extent follow principals of the Bismarck's model as gradual decentralization of management and financing have been implemented since the early 1990s. From economic and administrative viewpoints this decentralization is reflected by the liberalization of healthcare market, which results in gradual replacements of state health care units by municipal and non-public entities. Local governments (provinces, counties and municipalities) are now allowed to found and manage health care units what is permitted by the *Act of 30th September 1991 on Health Care Units*. Four forms of health care units are mentioned in this act with respect to their ownership and financial system: independent public health care centers (SPZOZ), budget entity, self-governmental budgetary establishment, and non-public health care unit (NZOZ). Decentralization of financial system and transformation of health care units into independent public health care centers began in 1998 and 1999 after sickness funds were established. This decentralization was reversed in 2003 as National Health Fund was founded, something that led to concentration of financial resources on a national level. The return to central healthcare financier and insurer was fiercely criticized by the politicians and scientific community; thus regional branches of NFZ were established, each responsible for healthcare financing and insurances in one province (but still operating under the Ministry of Health). Consequently, a lack of state-independent insurer and provider of healthcare limits patients' choice, which plainly contradicts the Bismarck's model principals. Contrarily to the healthcare financing and insuring, the responsibilities of management and planning in healthcare were imposed on local governments. Unfortunately, local communities were not able to cover increasing expenses and debts which health care units amassed over the years. These debts resulted from operation in accordance with constitutional principle of equal and free access to health services as well as life saving obligations. Moreover, the financial burdens are excavating prompted by the inability to declare bankruptcy by these health care units which are crucial for local population in order to retain overall health security intact.

4. Spatial inequalities in the availability of primary and hospital care

Legal and administrative characteristics concerning the organization of Polish healthcare system have a direct impact on spatial issues of the essence for researchers in medical geography. From an economic perspective, geographical sciences may lie beneath the premises for allocation of funds in particular regions (in accordance with spatially diverse needs) as well as the distribution of decisive and executive competences in health policy (spatial scale problem – consistency between administrative level and responsibility for health policy goals). From a social perspective, a key issue is to increase accessibility to and availability of medical services for all citizens, particularly the poorer friction of population who reside in peripheral areas.

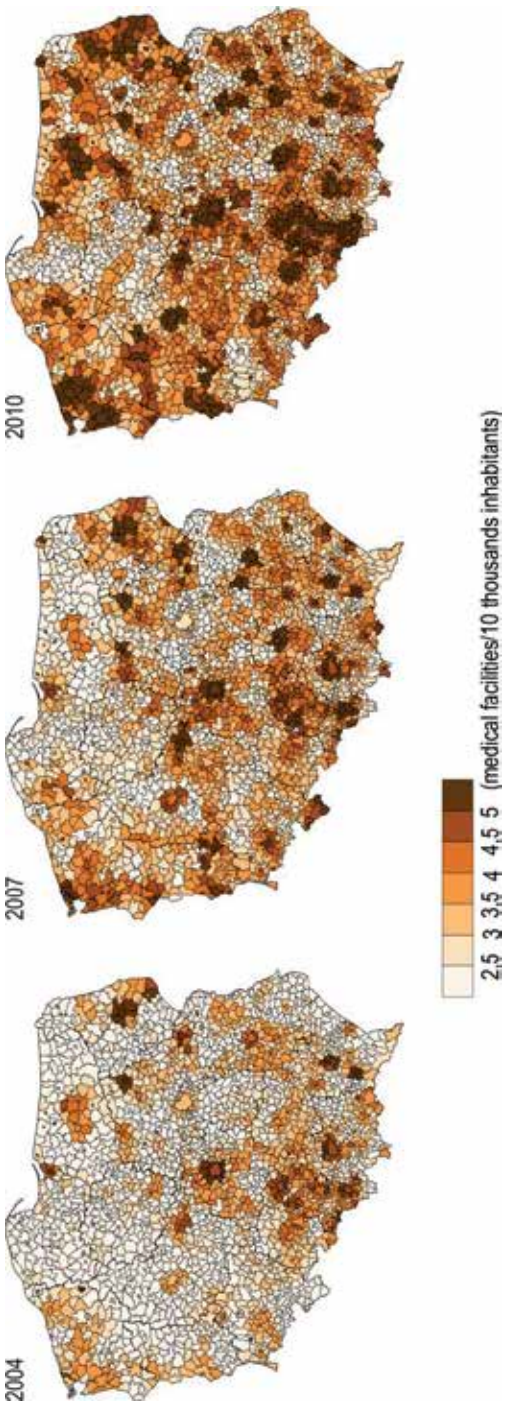
Both spatial accessibility and availability of healthcare was radically improved by the enforcement of legal acts that gave non-public entities rights to run healthcare practices as

well as make contracts with the National Health Fund. These novelties triggered a continuous increase in the number of health care units and health care centers across the country since 1991, but most intense spread was observed at the beginning of the 21th century. The data currently available allow for investigating these phenomena in a spatial dimension since 2004 (when a computer database – Register of Health Care Units was created). The number of health care units in 2004, 2007 and 2010 per 10 thousand inhabitants is presented in the Figure 2 irrespective of medical specialty and organizational forms. Higher number of health care units contributes to better availability and diversity of medical services. The changes in the number of health care units are inarguably connected with an increase of non-public entities. Because the range of influence for certain units often exceeds the municipal or regional borders (patients choose health professionals located nearby their places of residence) an average measure was calculated. This measure combines each municipality and all adjacent to them according to queen contiguity spatial weights frequently used in spatial statistics.

Most health care centers are located in the largest metropolitan areas both in cities and their vicinity as the suburban inhabitants often utilize health services provided by the institutions located in the inner city. More favorable healthcare availability in metropolitan areas results from large demographic potential, extensive financial resources, and excellent access to specialized medical services, the latter caused by high-rank education provided by medical universities that educate most qualified personnel. Thus, health professionals who obtained rare specializations usually practice in the largest cities. Furthermore, hospital wards with a catchment area encompassing several provinces (e.g. due to the uncommon specialization and rare disease treatment) are located in the largest cities too. Nevertheless, certain areas such as medium-sized towns, especially former province capitals (between 1975 and 1999 Poland was divided into 49 provinces; this period gave an economic boost to provincial capitals), are distinguished by a high-level and numerous medical services. The infrastructure inherited from the period of the People's Republic of Poland contributed to the concentration of health care units in these towns presently. Regional approach demonstrates Śląskie, Łódzkie, Zachodniopomorskie and Podlaskie Provinces as those of the best healthcare availability. Fast pace of changes can be observed in Zachodniopomorskie and Podlaskie Provinces whereas fewer health care units per 10 thousands inhabitants can be found in Kujawsko-Pomorskie, Pomorskie, and Warmińsko-Mazurskie Provinces.

4.1 Primary heathcare

General practitioner is considered as a key element of primary healthcare in Poland. Individual GP practices were established quite recently – in 1991. These doctors are supposed to perform *gatekeeper's* role that is to provide entrance to the whole healthcare system as a first institution patients refer to. As follows, this role assumes that initial patient-doctor contact begins at general practitioner office (in Poland, these physicians are called family doctors). Theoretically, family doctors ought to possess enough knowledge and experience to cure (or at least assist) the majority of diseases; however they are granted a wide range of administrative competences. Aside from prescriptions, they can issue referrals to other specialists or hospitals and for numerous medical examinations.



Source: Authors' own work based on Central Statistical Office (www.stat.gov.pl).

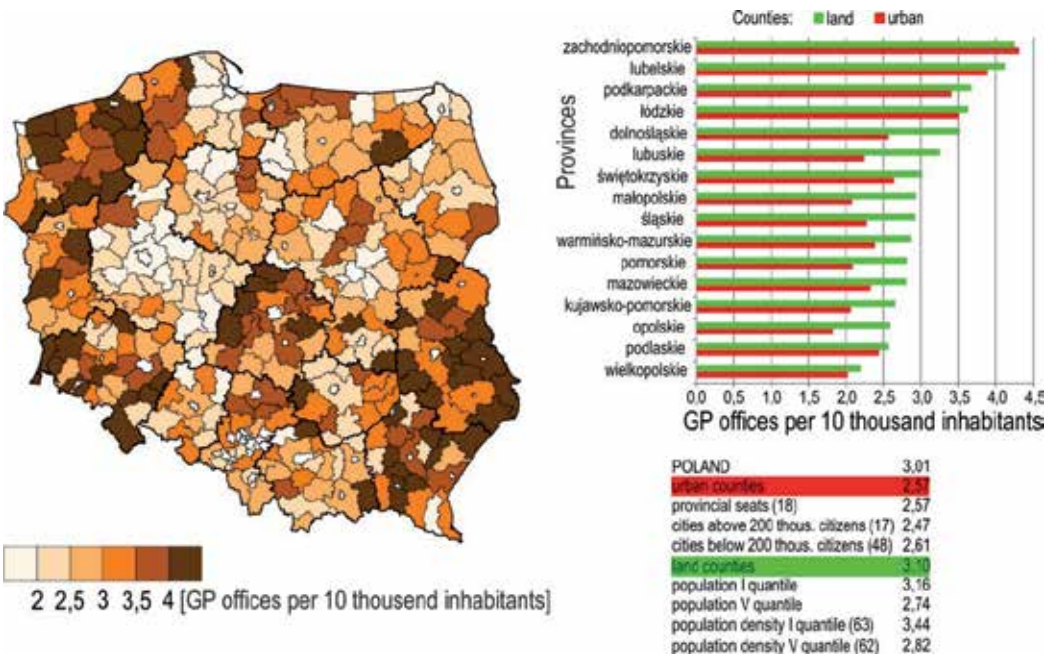
Fig. 2. The number of medical facilities per 10 thousands inhabitants in Poland in 2004, 2007 and 2010.

By this means, their medical functions are limited to distribution of prescriptions (most common and easy-to-cure diseases) or a set of referrals when the case is difficult to diagnose. Moreover, appointments aimed at receiving sick leaves are also very common. Therefore, the contemporary general practitioner office can be dubbed as a generator of referrals/hospitalizations and in this way whole sets of these documents can free family doctors from the responsibility to assist in more difficult and atypical cases. The original idea of family doctor assumes long-term and permanent contact between patients and health professional. The family doctor should be acquainted with patients' medical record and earn his trust through the years. However, frequent rotation of health professionals working in health care centers (mainly these located in cities) abides these goals from being obtainable. In sparsely populated rural areas, health services are usually provided by one general practitioner and a nurse. Moreover, in certain localities only branches of public health care centers are located and services are provided only during a few hours per day.

The foundation of healthcare model based on family doctor in Poland makes treatment more accessible and receivable everywhere, though the free-of-charge care can be obtained in both public and non-public health care unit, provided that the latter made a contract with the National Health Fund. The free choice of family doctor does not change the fact that the majority of patients are registered at general practitioner offices located in the vicinity of their places of residence. This solution is convenient for patients, especially when the change of residence in Poland does not require obligatory registration. Nevertheless, more changes of family doctor than twice a year requires from patient 80 PLN fee, unless this change is caused by a permanent migration to another place of residence, involuntary obligation, or results from other circumstantial conditions beyond patient's control. Patients who wish to change their primary care provider have to declare this will on a proper form. According to the National Health Fund a maximum number of patients registered to one general practitioner should not exceed 2750 (in other words: there should be at least 3.64 family doctors per 10 thousand inhabitants). However, the number of family doctors in Poland amounted to 10 206 in 2010, that is 2.67 family doctors per 10 thousand inhabitants (The Polish Chamber of Physicians and Dentists, www.nil.org.pl). This implies that there is rather a shortfall than excess of general practitioners in the country.

The distribution of primary health care units in Poland shows significant spatial diversity among provinces and counties (Figure 3). As for geographical factors of healthcare in different regions, this pattern does not directly refer to the level of socio-economic development, population distribution or historical background. Conversely, a noticeable diversity within certain regions can be observed. Among five provinces with the highest number of health care institutions per 10 thousand inhabitants there are regions of completely different background and socio-economic characteristics. Some diverse as far as primary healthcare availability is concerned counties are adjacent to each other. For example, they include counties located to the west of Poland (territories that used to be a part of Germany before World War II) in Zachodniopomorskie Province as well as less developed regions located to the east (so-called Eastern Wall) represented by Lubelskie Province. Similar diversity can be observed among provinces with the lowest availability of GP offices, e.g. Wielkopolskie and Podlaskie. Whereas the former can be considered as an area of economic prosperity, the latter is rather underdeveloped and experiences

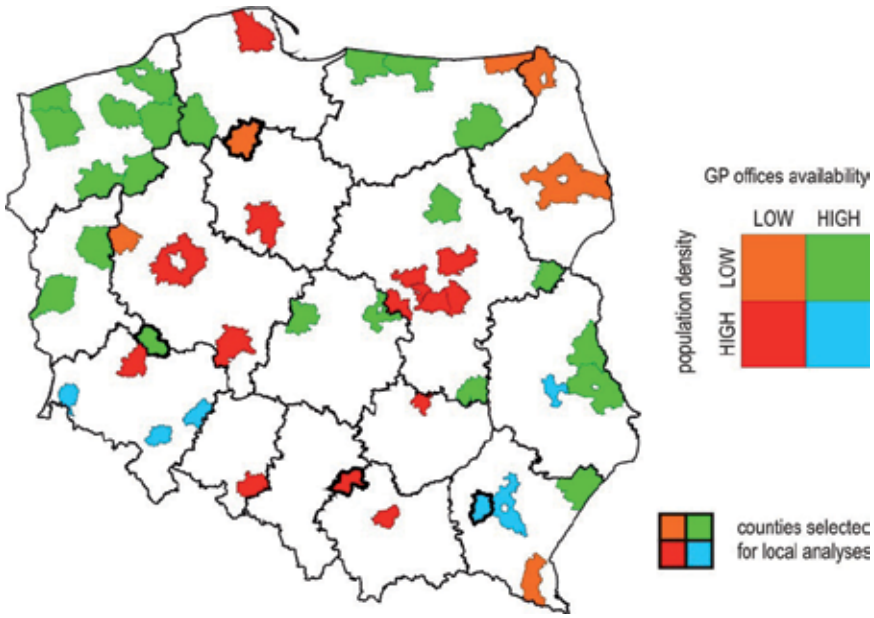
depopulation processes. Such diversity results from different models of healthcare organization and strategies implemented by local governments, but also spatial inequalities inherited after the Communistic times. In the areas of worse availability of primary care, family practices are probably larger as measured by the number of physicians in service. For this reason, significant differences can be observed between cities with county status (the largest towns and cities) and land counties. The majority of land counties are characterized by favorable accessibility of GP offices, what stems from higher population density and concentration medical facilities as a part of larger health care centers. As for land counties more GP offices per 10 thousand inhabitants are present in those with the lowest population number and density. Despite theoretically lower demand for medical services in these areas the network of GP offices is left uninterrupted what minimizes the distance between patient residence and family doctor.



Source: Authors' own work based on Register of Health Care Units (www.rejestrzoz.gov.pl) and Central Statistical Office data (www.stat.gov.pl).

Fig. 3. The number of general practitioner offices per 10 thousand inhabitants in Polish counties and provinces in 2010.

Analyses of primary healthcare on a regional level hide local disparities in health care accessibility reflected by the distribution of population. In order to detect conditions that underlie the availability of primary healthcare four types of areas (counties) in the whole country are singled out. The prerequisite for this selection is a simple spatial typology created according to the number of GP offices per 10 thousand inhabitants and population density (Figure 4).



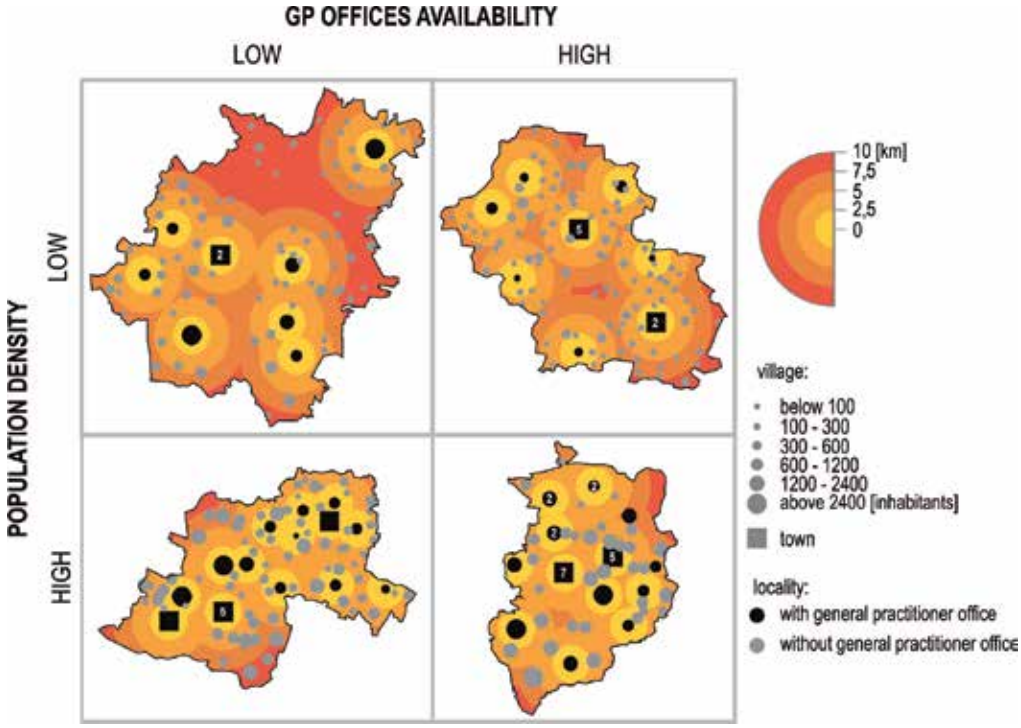
Source: Authors' own work based on Register of Health Care Units (www.rejestrzoz.gov.pl) and Central Statistical Office data (www.stat.gov.pl).

Fig. 4. The rural counties in I (low) and V (high) quintile in general practitioner availability as measured by offices per 10 thousand inhabitants or I (low) and V (high) quintile in population density in 2010 (encircled counties were selected for further local analyses).

The selected areas include:

a. Tucholski County (Kujawsko-Pomorskie Province – orange color) placed in the I quintile in GP offices availability and population density
b. Górowski County (Dolnośląskie Province – green color) placed in the V quintile in GP offices availability and I quintile of population density
c. Olkuski County (Małopolskie province – red color)placed in I quintile of GP offices availability and V quintile of population density
d. Ropczycko-Sędziszowski County (Podkarpackie Province – blue color) placed in V quintile of GP offices availability and V quintile of population density (Figure 5).

The example of Tucholski County shows that spatial accessibility to family doctors is constrained only in sparsely populated areas of the low GP offices availability. Within Tucholski County there are numerous small villages located further than 10 km from the nearest GP office, although most are located within the range of 3 km. Such areas are rather rare in Poland and can be found only in the northern part of the country and in some municipalities located in the Carpathians. Among the counties with high population density spatial accessibility is comparable for both the areas of low and high GP offices availability. Nonetheless, the high concentration of GP offices in towns elevates the indicator for the whole county (Ropczycko-Sędziszowski County). The areas characterized by high density of population and considerable number of GP offices are located only in southern Poland.



Source: Authors' own work.

Fig. 5. Primary healthcare in four counties of different population density and GP offices availability in 2010.

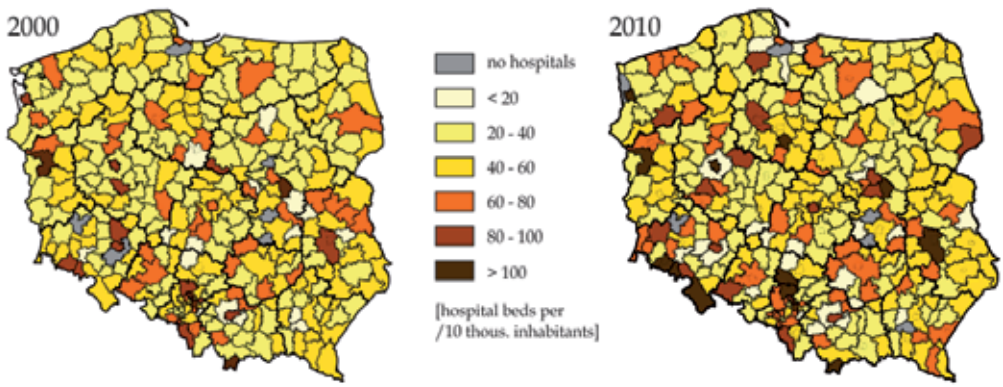
On the other hand, many sparsely populated counties have a high GP offices availability, what is particularly prominent in Zachodniopomorskie Province. The example of Górowski County demonstrates that, unless there are villages without family doctor offices in operation, the majority of such villages are located within 5 km distance from the nearest GP office. The opposite situation exists in Olkuski County, where larger villages are located relatively far from GP offices and some concentration of health care institutions is observed only in towns and adjoining villages. Interestingly, within the counties of low availability of GP offices and high population density fall some suburban areas of the biggest cities in Poland such as: Poznań, Warszawa, Gdańsk and Kraków. This proves that inhabitants of metropolitan areas utilize the healthcare services in the central city, what decreases the demand for GP offices in suburban areas.

4.2 Hospital healthcare

Inappropriate spatial and organizational structure of Independent Public Health Care Centers (SPZOZ) is believed to underpin the inequalities in Polish inpatient healthcare (Ruszkowski, 2008). Undoubtedly, higher actual net needs (in this case the number of hospitalized patients) concern large urban centers, what directly results from their demographic potential. However, healthcare needs are considered to be the best satisfied

not in urban centers with a high concentration of healthcare resources, but in sparsely populated areas, where there is one large hospital (Ruszkowski, 2010).

These inequalities in access to hospitals could have been mitigated alongside with the implementation of legislative Act on Network of Hospitals. The project of this legislation propounded a set of criteria to decide which institutions should be incorporated into Polish hospital network (hospitals that do not fulfill the criteria were either to be shut down or privatized). Among other things, these conditions included the optimal number of hospital beds with a regard to geographical distribution of medical resources. So-called regional adjustment plans were supposed to take into account "*the directions of hospital infrastructure development, demographic and epidemiologic determinants and their changes in time, the structure and length of hospitalizations*". Besides, these plans had to include the provision of sufficient accessibility to high-quality health services. As follows, according to the guidelines provided by the Ministry of Health a hospital must have at least 150 beds and the minimal number of beds per 10 thousand inhabitants should not be less than 40 (*The projected Act on Network of Hospitals, 2007*). Aforementioned project was vetoed in January 2009 due to political reasons and strong criticism from local governments (particularly controversial was the issue of closing down small hospitals). Though turned down, this project showed a great importance of geographic aspects concerning hospital network and its organization. No sooner than 15 years after transformation did the decision makers notice a need to fix the unfavorable distribution of tertiary healthcare in Poland.



Source: Authors' own work based on Healthcare Register (www.rejestrzoz.gov.pl) and Central Statistical Office data (www.stat.gov.pl).

Fig. 6. The number of hospital beds per 10 thousands inhabitants in 2000 and 2010 (without psychiatric hospitals and facilities located in health resorts).

Currently, an increase in the number of hospital beds Poland is observed, what was not the issue shortly after socio-economic transformation in the 1990s. Figure 6 shows some mosaic-like disparities across Polish counties as measured by hospital beds per 10 thousand inhabitants. In many counties this indicator falls below the recommended 40 beds per 10 thousand inhabitants. The regions "abundant" with hospital beds include Dolnośląskie and Śląskie Provinces. Urban areas, especially these of the largest Polish cities, possess relatively high number of hospital beds, which rarely fall below 60 beds per 10 thousand patients. This

surplus is utilized by the population of counties located in suburban zones where there are either no hospitals or some small unspecialized institutions. Counties located along the province boundaries have considerably lower number of hospital beds. Importantly, one large hospital, even though located in a small county, may have a broad catchment area. As a consequence, adjoining counties have fewer hospital beds per 10 thousand inhabitants. This attests to the inequality in spatial distribution of tertiary care in Poland. This problem particularly concerns large institutions, sizes of which often exceed local demand. Simultaneously, such hospital catchment areas become large and attract patients residing in more distant areas with no at all or only small general hospitals. The concentration of hospital resources in one place is perceived as profitable and socially approvable when these institutions offer a wide variety of specialized health services and operate as centers of scientific research and new technologies (Ferguson et al., 1997). The selective concentration of specialized hospital infrastructure in 1960s and 1970s in Poland resulted in too many hospital beds which cannot be explained neither by local demand nor accessibility of qualified personnel.

In 2000, there were 49.9 hospital beds per 10 thousand inhabitants in Poland. During next ten years this proportion increased to 55.2. However, the observed number did not always increase in accordance with in the improvements in accessibility of stationary health care and across particular medical specializations. The research conducted by the Centre of Health Care Organization and Economics by the end of 1990s showed significant inequalities in spatial distribution of long-term care beds and the necessity to increase their number significantly (Kozierkiewicz, 2008). Results of the study conducted by the National Institute of Hygiene indicated that the greatest excess of hospital beds concerns such wards as: ophthalmology, otolaryngology, pediatric surgery, obstetrics and gynecology, and especially in the west of Poland. On the other hand, shortages of hospital beds can be found on rehabilitation, hematological and oncological wards (Goryński et al., 2006). Aforementioned study was questioned by Murkowski (2007), who argues that the largest surplus of hospital beds is observed in Śląskie, Łódzkie, Lubelskie and Podlaskie Provinces, so in the central and eastern Poland. This finding is more or less congruent with the results presented in the Figure 6. On the other hand, reflections presented by Krzanowski (2007) are somewhat controversial and reveal alleged influence of healthcare system on hospital morbidity. As pointed out by this author, regional differences in hospitalization rates for certain diseases are well explained by the available number of hospital beds. Medical geographers with a sufficient experience and skills in finding spatial relations between the needs and supplies in many socio-economic domains should be included into researchers exploring this phenomenon. The application of causative and consecutive analyzes would help find solutions for Krzanowski's concern that is to tell whether and in which regions the statement "*if there are spare beds, there will also be patients*" can be true.

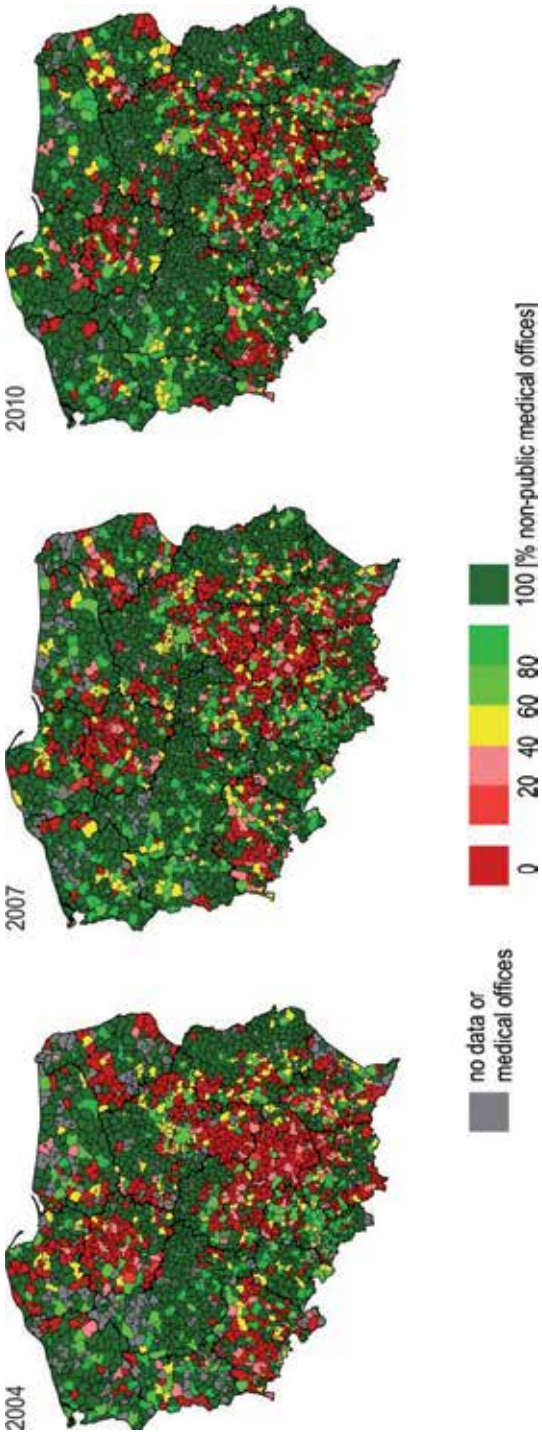
5. Commercialization and privatization in Polish healthcare

Commercialization and privatization in Polish healthcare are considered as key determinants of spatial and non-spatial availability to primary and hospital care in the recent years. From a spatial perspective these processes lead to an increase in the number of medical facilities, however commercialized healthcare limit affordability for both insured and uninsured citizens as some services are paid. Commercialization does not occur

uniformly throughout the country and across medical specialties. In 2009, about 75% general practitioner offices (max. Wielkopolskie 94%, min. Świętokrzyskie 48%), 82% specialist offices (Wielkopolskie 91%, Świętokrzyskie 70%) and 45% of general and specialized hospital facilities (Dolnośląskie 68%, Świętokrzyskie 24%) belonged to commercial entities. Changes in ownership structure are clearly reflected by *inverse care law* – a concept developed by Hart in 1971 (Hart, 1971). This law assumes that the availability of good medical care tends to vary inversely with the need of the population served. In other words, financial resources are not allocated in conjunction with the distribution of needs, but rather along with the distribution of resources. Location and quality of health services offered by non-public health care units become market-oriented and favor more affluent regions and social groups. The poorer patients are not as attractive customers as other inhabitants despite of higher needs reported by the former. Such situation is most characteristic for the USA – a country with a dominance of private healthcare financed by non-public insurance companies (except for the elderly and low-income groups). This organization of healthcare Whiteis (1997) calls „*corporate-sponsored medicine*".

Spatial aspects concerning privatization and directions of ownership changes are presented on the example of all medical facilities, general practitioner offices and hospitals. Figure 7 depicts transformations of public health care units into non-public entities for 2004, 2007 and 2010. According to the *Act of 30th September 1991 on Health Care Units* non-public health care units can be founded by: churches and religious groups, employers, foundations, trade unions, professional or other associations, other national or foreign legal or natural persons or non-legal partnerships.

Changes in health care unit ownership structure evidenced by an increase of non-public facilities progresses rapidly in the whole country. In small rural communities private entities get complete or partial hold of municipal health centers followed by a contract drew with the National Health Fund. As for primary care, almost all services remain refunded by the NFZ, but in case of specialist offices some services are paid. Thus, in many areas the spatial accessibility of healthcare increases as branch offices are more likely to be opened by commercialized health care units, but this happens selectively (usually in the largest villages). Private medical offices in large cities remain market-oriented and operate under great competition. Therefore, firms locate their offices is strategic locations usually in the vicinity of potential clients e.g. in large shopping centers. A good example of this is Enel-Med healthcare provider which possesses offices in the biggest shopping malls across the country: Arkadia and Blue City (Warsaw), Galeria Krakowska (Cracow), Arkady Wrocławskie (Wrocław), Manufaktura (Łódź) and Kupiec Poznański (Poznań). Some of the services offered by private healthcare firms are not refunded by the NFZ, so, in spite of improved spatial accessibility, their affordability is limited. In less populous urban areas and towns medical offices are often located in private houses what is rare in the bigger cities (except for dental offices). Unfortunately, in spite of rising market-oriented availability the possibilities to utilize health services are constrained by too high demand and annual limitations for certain services and their refund by the NFZ. As a consequence, long lines and wait times to the physicians are observed what discourages the patients and attracts them to utilize paid services (frequent in specialist offices). Free healthcare can be utilized without a need to wait after private (and paid) consultation – such practices are not uncommon.

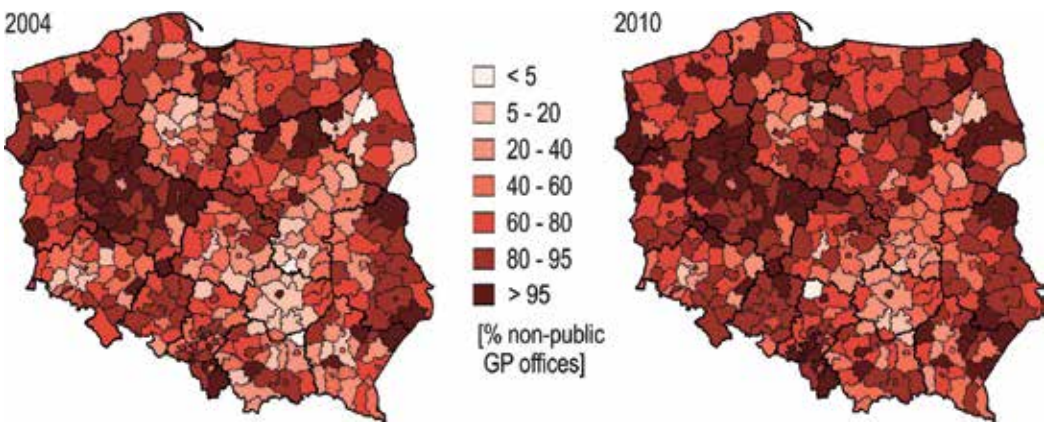


Source: Author's own work based on Central Statistical Office (www.stat.gov.pl)

Fig. 7. The share of public and commercial healthcare facilities in 2004, 2007 and 2010.

In 2004, public healthcare units prevailed only in Świętokrzyskie and Kujawsko-Pomorskie Provinces, but in 2010 the majority of healthcare facilities in the whole country were non-public. Commercialization in the Polish healthcare first dominated Wielkopolskie Province, where liberal and entrepreneurial attitudes prevail among the local population. Conversely, in Świętokrzyskie and Mazowieckie Provinces left-wing political parties traditionally gain great popularity among traditionally pretentious communities. These parties strive to delay the privatization of healthcare in fear of paid services and undermined health security. Such social attitudes are clearly reflected in election results; therefore more conservative municipal authorities are not likely to foster quick changes in health care unit ownership. This selective commercialization depends on leading political fraction in local governments. Perhaps, low availability and quality of services in some areas make their residents press on local authorities for non-public care irrespective off political affiliations. This factor may be of the essence in the eastern part of the country

In primary healthcare commercialization processes occur very fast (Figure 8). In 2004 most counties had more than a half of their general practitioner offices commercialized. In 2010 about 75% of all GP practices belonged to non-public entities. Three separate areas of sizeable prevalence in the proportion of non-public facilities can be demarcated: prosperous and liberal west (Lubuskie and Wielkopolskie Provinces), most industrialized south of population traditionally emigrating to the west, Germany in particular (Śląskie and Opolskie) as well as poor eastern borderland (Lubelskie). On the other hand, Świętokrzyskie Province, the south of Mazowieckie and Łódzkie Provinces are represented by a dominance of people's and left-wing political affiliations; thus comprise a majority of areas with public healthcare offices. In other regions decisions against commercialization depend on specific local determinants.



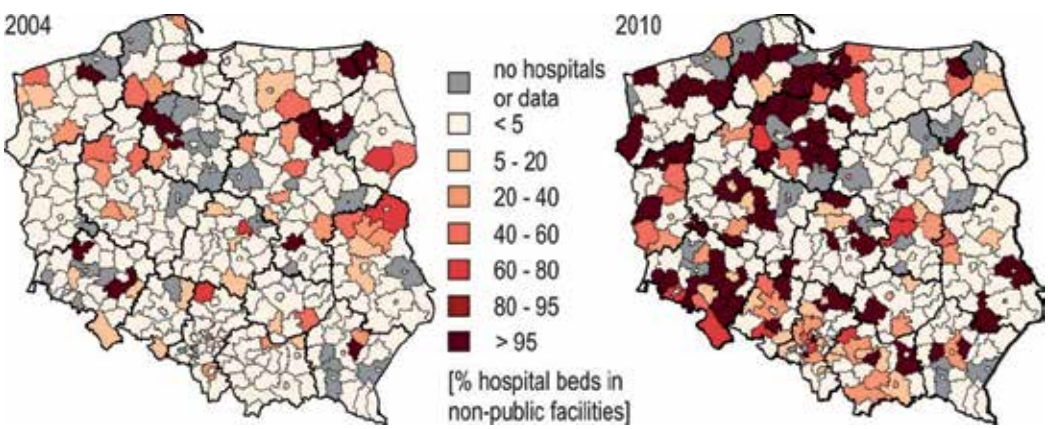
Source: Authors' own work based on Register of Health Care Units (www.rejestrzoz.gov.pl).

Fig. 8. The share of non-public general practitioner offices in 2004 and 2010

According to some scientific views the most financial problems experienced by hospitals can be easily solved by changing their ownership to join stock or commercial companies (Masiakowski, 2005; Milczarek, 2005; Rój, 2006). This does not mean that the majority of independent public health care centers (mainly hospitals) need to be converted into private

properties because they operate with no regard to economic rationality (Jończyk, 2008). Unfortunately, considerable service overproduction and capital intensity (rising needs, new technologies in medicine and pharmacy) combined with simultaneous financial scarcities in municipal budgets led to large indebtedness. This is the main reason why commercialization processes are at issue and worry local and regional authorities as well as politicians. However, ownership changes do not result in a complete lack of control over hospitals as most shares are often hold by public bodies; thus, instead of privatization, the term commercialization better illustrates current transformation in the Polish healthcare (Misińska & Nawara, 2008). These processes are selective and connected with restructuring and reforms in healthcare system, but particularly with vigorous attempts to clear hospitals of liability for debts. The germ of these endeavors was to be *the Act of Regulations of Healthcare Legislations* (so-called healthcare legislation package), which proposed mandatory choice: to convert all public hospitals into commercial companies or pay their debts by public owners (local and regional governments). In 2009, this legislation was vetoed by President Lech Kaczyński, who refused to allow for paid services and put population health security in jeopardy as some unprofitable hospitals could have been shut down in the aftermath of the new code.

Nevertheless, commercialization of healthcare facilities continues, partially fueled by financial aid of so-called *Governmental B Plan*. This plan promises a donation for these governments which commercialize hospital, make over all assets, property and other resources (unless a new owner already possess the resources necessary to run a hospital), and designate the entity that would take over all of the debts amassed. Most frequently, commercialized independent health care centers are converted to limited liability companies with shares held by local or regional governments. In 2004, 6.4% of all hospital beds were owned by non-public entities. In 2010, this indicator increased to 15.5%, so 32.8 thousand hospital beds per 210.8 thousand in the whole Poland belonged to non-public bodies. The spatial depiction of these transformations is presented in the Figure 9.



Source: Authors' own work based on Register of Health Care Units (www.rejestrzoz.gov.pl).

Fig. 9. The share of hospital beds in non-public facilities in 2004 and 2010 (without psychiatric and facilities located in health resorts, urban counties are combined with rural except the largest cities).

The largest proportion of commercialized hospitals between 2004 and 2010 concerned Kujawsko-Pomorskie, Pomorskie and Dolnośląskie Provinces. Usually, one or largest facility was commercialized in one county, thus small disparities within counties contrast with large disparities between them. The latter, though, show a mosaic spatial configuration as a direct consequence of selective processes dependent on either undisputed and quick decisions or social protests and political unwillingness among the decision makers. Nevertheless, faster changed can be observed in western Poland what is comparable to the level of socio-economic development.

6. Conclusions

In this study Polish healthcare system is characterized from spatial and organizational viewpoints. The analyses conducted show considerable regional and local disparities in access to health services across the country. While existing inequalities are evident and allow for delineation of excess/shortage areas as far as health resources are concerned, the evolution of geographical studies should aim at seeking spatial relationships between healthcare resources and population needs. Such approach has been put into practice in the USA health policy. The main purpose of Health Professional Shortage Areas (HPSA) is to identify areas of greater need for health care services and redirect limited healthcare professional resources to people in those areas. This objective is congruent with the geography of healthcare principal that is matching healthcare resources to population needs in time and space. Consequently, some apparent scientific goals also come to the fore. These include more complex cross comparisons between volume and structure of health resources and volume and structure of population needs. Such multivariate analyses ought to be able to provide answers to simple questions i.e. how many hospital beds do we need?; What hospital wards need to be expanded or downsized?; What is the optimal number of general practitioners/ practitioner offices per 10 thousand inhabitants in an area including current health, demographic and economic situation?.

Limited geographic access to primary care in Poland concerns only areas of very low population density, which are not as common as in other European counties. For that reason non-spatial limitations (including financial and legal) should forge ahead when analyzing utilization of healthcare facilities (Jones & Moon, 1987; Powell, 1995).

This study introduces to the complexities of Polish healthcare legal and administrative foundations and spatial availability with a special regard to difficulties adversely affecting patients' access to healthcare facilities. To conclude, spatial and organizational availability of healthcare in Poland are shaped by the following phenomena and processes:

a. ineffective and extensive management of healthcare resources during the communistic times, lack of regional and local health policies, system centralization (state organization, management, planning and financing)
b. hospital locations in 1945-1989 unrelated to population needs
c. transition from budgetary to insurance healthcare model (similar to Bismarck model), contract-based financing of healthcare
d. increasing number of health care units → higher accessibility and availability (much better in metropolitan areas)
e. mosaic spatial distribution of hospital care (better availability explained by the proximity to large general hospital)

f. increasing needs along with advancements in medicine and pharmacy as well as better availability of healthcare facilities
g. various quality of hospital services linked to the following dilemma: does contemporary hospital heal or perform contracts? A quest for balance between money saving, debt reduction and patient needs
h. the role of the general practitioner: limited in treatment, but major in administration (sick leaves, prescriptions, referrals); organizational and spatial access impeded in case of secondary healthcare
i. threats of hospital privatization and healthcare security of citizens, declining role of small hospitals, even though some may be important for peripheral areas
j. faster ownership change in the west of Poland parallel to rising socio-economic development and liberal attitudes among the locals

7. References

Chawla M., Berman P., Windak A., Kulis M., *Provision of ambulatory health services in Poland: a case study from Krakow*, Social Science & Medicine, 58, pp. 227-235.

Curtis S., Malczewski J., 1990, *Planowanie przestrzennej alokacji wydatków na ochronę zdrowia w Anglii i w Polsce - zarys badań porównawczych*, [w:] Smoleń M. (Ed.), *Teoria i praktyka organizacji ochrony zdrowia. Przestrzenne planowanie finansowe opieki zdrowotnej. Elementy teorii. Próba praktycznych rozwiązań*, Instytut Medycyny Pracy, Łódź, pp. 63-81.

Dziubińska-Michalewicz M., 1994, *Sektor prywatny w opiece zdrowotnej, wyniki badań ankietowych*, Antidotum - Zarządzanie w Opiece Zdrowotnej, 12, pp. 27-34.

Ferguson B., Sheldon T.A., Posnett J., (eds.), 1997, *Concentration and Choice in Healthcare*, FT Healthcare, London.

Jones D. R., Moon G., 1987, *Health, Disease and Society*, London, Routledge and Kegan Paul.

Jończyk J., 2008, Aspekty *prywatyzacji szpitala*, Praca i Zabezpieczenie Społeczne, 7, pp. 2-7.

Joseph, A.E., Phillips, D.R., *Accessibility and Utilization - Geographical Perspectives on Health Care Delivery*, Happer & Row Publishers, New York, 1984.

Hart, J., *The Inverse Care Law*, The Lancet, 297, 1971, pp. 405-412.

Haynes R., Lovett A., Sunnenberg G., 2003, *Potential accessibility, travel time and consumer choice: geographical variations in general medical practice registration in Eastern England*, Environment and Planning A, Vol. 35, pp. 1733-1750.

Goryński P., Wojtyniak B., Kuszewski K., 2006, *Ile potrzeba nam łóżek szpitalnych – załącznik do projektu ustawy o sieci szpitali*, Menedżer Zdrowia 9/2006.

Grochowski M., 1988, *Rejonizacja służby zdrowia a dostępność usług medycznych*, Rozwój Regionalny, Rozwój Lokalny, Samorząd Terytorialny 15, Uniwersytet Warszawski, Wydział Geografii i Studiów Regionalnych, Warszawa.

Guagliardo M.F., 2004, *Spatial accessibility of primary care: concepts, methods and Challenges*, International Journal of Health Geographies 3, 3. pp. 1-13.

Kaczmarek T., Marcinkowski J. T., Zysnarska M., Maksymiuk T., Majewicz A., 2007, *Nierówności społeczne w dostępie do zdrowia*, Problemy Higieny i Epidemiologii, 88, 3, pp. 259-266.

Kaser M., 1976, *Health care in the Soviet Union and Western Europe*, Croom Helm Limited, London.

Kearns R. Moon G., 2002, *From medical to health geography: novelty, place and theory after a decade of change*, Progress of Human Geography, 26, 5, pp. 605-625.

Kozierkiewicz A., 2008, *Koło ratunkowe dla szpitali. Od doświadczeń do modelu restrukturyzacji*, Termedia, Warszawa.

Krzanowski M., 2007, *Są łóżka, będą i chorzy…*, Rynek Zdrowia 4/2007, pp. 48-49.

Malczewski J., 1989, *Optymalizacja obszarów obsługi placówek podstawowej ochrony zdrowia*, Przegląd Geograficzny, T. LXI, z. 1-2, pp. 23-31.

Masiakowski A., 2005, *Prywatyzacja w ochronie zdrowia*, Zdrowie Publiczne 115, 2, pp. 252-253.

Mayer, J. D., 1982, *Relation between two traditions of medical geography: health system planning and geographical epidemiology*, Progress in Human Geography, 6, pp. 216-230.

Mazurkiewicz L., 1994, *Czy geografia człowieka powinna zajmować się problematyką zdrowia*, Przegląd Geograficzny, t. LXVI, z. 1-2, pp. 191-195.

Michalski T., 1999, *Nowe nurty w światowej i polskiej geografii medycznej*, Kwartalnik Geograficzny, 4, 12, pp. 85–89.

Milczarek M., 2006, *Warunki ekonomiczno-finansowe działalności i rozwoju szpitali. Perspektywy i niezbędne działania*, Polityka Zdrowotna, t. 3, Instytut Polityki Ochrony Zdrowia przy Uniwersytecie Medycznym, pp. 7-9.

Millard F., 1995*, Changes in the health care system in post-communist Poland*, Health & Place 1, 3, pp. 179-188.

Misińska B., Nawara P., 2008, *Publiczna i prywatna własność w systemie ochrony zdrowia w kontekście form organizacyjno-prawnych prowadzenia działalności medycznej*, [w:] Ryć K., Skrzypczak Z. (Ed.), *Ochrona zdrowia i gospodarka – mechanizmy rynkowe a regulacje publiczne*, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa, pp. 335-344.

Ministerstwo Zdrowia, 2006, *Wskaźniki do tworzenia projektu tworzenia sieci szpitali z elementami analizy sytuacji demograficznej i stanu zdrowia ludności*, Materiał przygotowany dla Ministra Zdrowia przez Państwowy Zakład Higieny na podstawie danych Centrum Informacyjnych Ochrony Zdrowia i Państwowego Zakładu Higieny dotyczących infrastruktury szpitali i ich działalności, Warszawa.

Moon G., Gould M. and Jones K., 1998, *Seven up – refreshing medical geography: an introduction to selected papers from the Seventh International Symposium in Medical Geography*, Portsmouth, UK. Social Science & Medicine 46, pp. 627–30.

Murkowski M., 2007, *W sieci niekompetencji*, Menedżer Zdrowia, 4/2007, pp. 21-24.

Parr, H., 2003, *Medical geography: Care and caring*, Progress in Human Geography, 27, 2, pp. 212-221.

The projected act of Network of Hospitals, (Projekt ustawy o krajowej sieci szpitali), Ministerstwo Zdrowia, available on the Ministry of Health website website:http://www.mz.gov.pl/wwwfiles/ma_struktura/docs/u_siec_szpitali_kerm_17052007.pdf, Access: 12/17/2010.

Penchansky R., Thomas J.W., 1981, *The Concept of Access*, Medical Care 19, 2, pp. 127-140.

Powell M., 1995, *On the outside looking in: medical geography, medical geographers, and access to health care*, Health&Place, Vol. 1, No. 1., pp. 41-50.

Register of Healthcare Units official website – www.rejestrzoz.gov.pl.

Rosenberg M. W., *Medical or health geography?: populations, peoples and places*, International Journal of Population Geography, 4, 1998, pp. 211-226

Rój J., 2006, *Forma organizacyjno-prawna a gospodarka finansowa szpitala*, [w:] Węgrzyn M., Wasilewski D. (Ed.), *Komercjalizacja i prywatyzacja ZOZ - kluczowe warunki osiągnięcia sukcesu*, Prace naukowe AE Wrocław, pp. 53-58.

Ruszkowski J, 2008, *Polski system zdrowotny – socjalizm w rynkowym otoczeniu*, [w:] Ryć K., Skrzypczak Z. (Ed.), *Ochrona zdrowia i gospodarka – mechanizmy rynkowe a regulacje publiczne*, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa, pp. 29-42.

Ruszkowski J., 2010, *Zwiększenie bezpieczeństwa zdrowotnego*, ekspertyza finansowana ze środków projektu nr POPT.03.04.00-00-019/07 w ramach Programu Operacyjnego Pomoc Techniczna 2007-2013 wykonana na zlecenie Ministerstwa Rozwoju Regionalnego.

Siwińska V., Brożyniak J., Iłżecka I., Jarosz M. J., Orzeł Z., 2008, *Modele systemów opieki zdrowotnej w Polsce i wybranych państwach europejskich*, Zdrowie Publiczne, 118, 3, pp. 358-367.

Świadczenia opieki zdrowotnej finansowane ze środków publicznych, Vademecum issued by National Health Fund, National Health Fund, Poland, 2011.

Unal E., Chen S.E., Waldorf B.S., 2007, *Spatial accessibility of health care in Indiana*, Purdue University Working Papers, West Lafayette, paper nr 07-07.

The Act of 28th October 1948 on Collective Health Care Centers and Planned Economy in Healthcare as published in Dziennik Ustaw No. 55, Item 434 (Ustawa z dnia 28 października 1948 r. o zakładach społecznych służby zdrowia i planowej gospodarce w służbie zdrowia, Dz.U. 1948 Nr 55, Poz. 434.)

The Act of 30th September 1991 on Health Care Units as published in *Dziennik Ustaw* No. 220, Item 1600 (*Ustawa z dnia 30 września 1991 o zakladach opieki zdrowotnej*; Dz.U. 1991, Nr 220, Poz. 1600.).

The Act of 6 th November 2008 on Code Indroducing Healthcare Legislations, Presidential veto - legislation not passed by the parliament (Ustawa z dnia 6 listopada 2008 r. Przepisy wprowadzające ustawy z zakresu ochrony zdrowia),
(http://orka.sejm.gov.pl/proc6.nsf/ustawy/294_u.htm), Access: 03/14/2011.

The Act of Regulations of Healthcare Legislations, vetoed by President Lech Kaczyński in 2009.

The Constitution Of The Republic Of Poland of 2nd April, 1997 as published in Dziennik Ustaw No. 78, Item 483.

The Polish Chamber of Physicians and Dentists official website - www.nil.org.pl

The Central Statistical Office Local Databank website - www.stat.gov.pl

Wang, F. and Luo, W., 2005, *Assessing spatial and nonspatial factors for healthcare access: towards an integrated approach to defining health professional shortage areas*, Health & Place, 11, pp. 131-146.

Whiteis D. G., 1997, *Unhealthy cities: corporate medicine, community economic underdevelopment and public health*, International Journal of Health Services, 27, pp. 227-242.

13

Planning Incorporation of Health Technology into Public Health Center

Francisco de Assis S. Santos and Renato Garcia
Biomedical Engineering Institute,
Federal University of Santa Catarina, Florianópolis,
Brazil

1. Introduction

The incorporation process of Health Technology (HT), particularly, Medical Equipment(s) (ME) encompasses all activities ranging from purchasing, renting, leasing or exchanging, technology assessment, planning and identification of needs, installation, technical rehearsals, calibration, users' training etc. The incorporation process also includes prediction of technology use for ascertaining if what has been planned can be realized, and for aiding future incorporations (World Health Organization [WHO], 2011a).

According to Wang (2009), the incorporation process of ME can be divided into two phases: planning and acquisition. The planning phase includes assessment of needs and impacts, and costs and benefits of ME after auditing the existing resources. The data collected during auditing and assessment should be established and converted into a technology incorporation plan, which might guide future investments. The second phase relies on the selection and acquisition of products that are appropriate to a certain application and environment. Purchasing options, such as leasing, lending and the revenue sharing models, should always be considered.

Health systems must be built in blocks in order to inform the financing policies, human resources, information, service aid, management and health technology. The inter-relations and interactions among these blocks constitute a system. If any of these is lacking, the health system cannot work on the level needed to improve public health. Each block has its own organizational and political challenges. This chapter will discuss the health technology block, considering ME as the essential tool to public health (WHO, 2007; WHO, 2009).

Technology in health service aid is indispensable, even in the most remote and low-resource areas. Drugs, implants, disposable products and medical equipment are the main items that contributed to the progress of health care in the last century, as compared with that during the preceding thousands years. Unfortunately, technology also adds significantly to the fast and ever-growing health costs. Within this context the ME stand for relevant costs to the health system and sometimes under low and limited resources, besides of many medical procedures being totally dependable of technological resources.

Management and administration of this health system technology, which aims at improving the cost-benefit ratio, safety, and reliability, falls within the domain of Biomedical Engineering. Clinical Engineering, which forms part of this domain, incorporates the quality parameters in all phases of the technology life cycle (Raymond, 2004; Moraes & Garcia, 2007).

Therefore, the Clinical Engineer, through ME management and administration, must identify the needs, limitations and factors required to evolve a methodology that leads to appropriate planning of ME incorporation through a systematized and rational structure. Thus, the health system can recommend incorporating just safe and effective ME that has infrastructure, human resources and financial viability. Moreover, it has to observe the legal, social and ethical aspects of the context in which the ME is to be inserted (Centers for Medicare and Medicaid Services, 2000; Cutler & Mcclellan, 2001; Sônego, 2007; Santos & Garcia, 2010).

Inadequate planning of ME incorporation practices can lower the quality of service aid or of ME's performance. On the other hand, adequate planning can lead to safe, equitable and quality health care. Besides, it also helps in identifying the technology that is appropriate to the Health Care Center (HCC)— not just the cheapest one taken from proposal selection (public bidding)—in terms of well defined and satisfactory parameters, such as deliverance, installation, performance test , training, payment and guarantee. Also, the technology must be so chosen as to encourage the distributors and manufacturers come back with future offerings (Calil, 2007; WHO, 2011a).

These guidelines are to be followed not only in case of purchases, but also in case of the equipments received through donation, renting or borrowing, including the ones replacing the existing ones. Moreover, should be applied to the individual institutions and/or network systems composed of several hospitals in various levels, health centers and community clinics, although the complexity and deadlines are very different from one case to another (Wang, 2009).

This chapter deals with identifying and recommending the main factors that must be considered for ME incorporation. The Clinical Engineer can help the actors involved, as a process facilitator, in identifying these factors and in deciding if incorporation is a real necessity. Thus, the performance of the Clinical Engineer strengthens not just the ME incorporation, but the whole health system and thus the public health.

2. Incorporation process of medical equipment

The main target of ME incorporation process is to maximize the benefits–clinical or financial—and minimize the costs—investment or recurrent ones-- especially of the local low resource communities, thus helping them in controlling the health problems effectively. The objectives may vary from one HCC to the other, but they usually include some of the following (Kaur, 2005a; Wang, 2009; WHO, 2011a; Santos & Garcia, 2010):

- Improve clinical results and patient satisfaction.
- Guarantee better access, quality and use of ME.
- Increase patients' life expectancy.
- Decrease the time spent in investigation, treatment and rehabilitation.

- Increase the access of patients to health care in equitable manner.
- Enlarge the coverage of patients' population and geographic areas.
- Reduce risks to patients, clinicians and environment.
- When suitable, keep or improve the ME market.
- Obtain balance between clinical needs, personal desire and available financial resources.
- Introduce pro-active planning to meet long-term needs, and thereby reduce emergency acquisitions.
- Reduce the Total Cost of Ownership (TCO).
- Offer more learning opportunities to clinicians and students when they are academically affiliated.
- Maintain or increase standardization to improve efficiency and reduce risks.
- Increase transparency of the public lender process.
- Encourage the actors involved in the incorporation process to create conditions conducive to establishment of monitoring actions towards the long life cycle of ME, and thus contribute to future planning.
- Observe the valid legal aspects in national and regional contexts.
- Identify the cultural and social barriers and facilitators.

It is important to note that other objectives can be added to the foregoing list depending on the need of each ME or health care. The Clinical Engineer must help in identifying the key factors for achieving the objectives defined. These factors must consider aspects inherent to technology, infrastructure, human resources and costs. They thus have a wide scope for choosing the parameters that meet the challenge of ME incorporation by using Clinical Engineering methodology.

After identifying the parameters, it will be possible to develop a systematized methodology that is based on the decision making domain, health technology assessment (HTA) and health technology incorporation. The Clinical Engineer can, therefore, act as a facilitator and as an actor of a team or interdisciplinary commission that formulates recommendations and supports decision making for ME incorporation, based on the evidence available in literature, in such a way as to minimize or even eliminate subjectivity in decision making.

2.1 Conceptual approach

A conceptual approach is needed to understand health technology, especially ME, its role and life cycle, and the actors involved in its incorporation process.

2.1.1 Medical equipment function

Health care is a human right, according to the Universal Declaration of Human Rights. However, it does not give access to universal health care. The World Health Report commented on this issue, in the context of primary health care, thus: "*Primary care and social protection reforms depend on choosing health-systems policies, such as those related to essential drugs, technology, human resources and financing, which are supportive of the reforms and promote equity and people-centred care*" (WHO, 2008; United Nations, 2011).

In this regard, it can be observed that health systems depend on health technology for the desired health results. It is important to plan ME programs according to the protocols and policies that can result in equitable, safe, appropriate and high technology access. ME requires adjustment, maintenance, repairing, user training and deactivation, which are usually performed by Clinical Engineers. ME is used for diagnostic purposes, treatment of certain diseases and rehabilitation with some kind of accessory input or other equipment. ME does not include implants and disposables (WHO, 2011b).

Technology[1] by itself has low intrinsic value and its value depends on how it is used. It is through the ME that health-predicted needs and benefits are realized, considering its impact on the patients, users, infrastructure, maintenance, costs and valid legislation. If the incorporation is planned and properly guided, then the ME can help policy formulators, decision makers, Clinical Engineers and health professionals in fulfilling their objectives of treating the patients under a better cost-benefit relation. However, if the technology is inappropriately incorporated or used, it can harm people, and cause loss of value and resources (Wang, 2009; National Institute for Health Research [NHS], 2010).

In this context, efforts must be made to manage the ME in a rational way, so that some balance can be found between the desired needs and benefits on the one hand, and the positive or negative impacts on the other. The importance of Clinical Engineering structures is thus evident in offering mechanisms that enable efficient and transparent ME incorporation planning. Nonetheless, the actors involved in the process must be aware that the incorporation is directly linked to the necessity of treating or diagnosing some clinical condition. Consequently, the eligibility of the applicant ME for incorporation must be assessed.

The eligibility refers to justification in realizing the ME assessment in the incorporation. To help this, some issues must be addressed, considering different aspects of demographic density, complexity of the health problem, and the nonexistence of unused ME in the HCC. This could enable the manufacturer and the distributor to guarantee the supply of spare parts during servicing of the equipment.

2.1.2 The medical equipment life cycle

ME is vital to health care service in that it improves the public health system. From the innovation phase to the replacement one, the tools used in the system must have four essential characteristics: availability; accessibility; adjustment; and financial capacity[2]. These would help to enhance the life cycle of the ME in such a way that not all the efforts may not have to be centered on the innovation phase alone, but on the incorporation one too in an adequate and rational manner; this ensures their use in an efficient and equitable way (WHO, 2011b).

---

[1]The majority of dictionaries define technology as the application of knowledge to practical means.

[2]Relationship between prices of services according to the maintainer, the deposit required for the entry of customers and the ability to pay, or the existence of health insurance by the customer.

In general, ME life cycle presents four phases (WHO, 2011b):

- Research and Development (R & D).
- Regulation.
- Health Technology Assessment (HTA).
- Health Technology Management (HTM).

These phases are efficient as long as they are supported by the health policies which are supervised by trained personnel. While interdependence of these phases is important to achieve the desired results, the operation within each phase must also be planned and executed with protocols that correspond to the administrative level (national, regional and local) (WHO, 2011b).

In the R&D phase, the entry parameters depend on the national policy of health technology R & D and on the health needs of the population. Besides meeting these requirements, the national policy must concentrate on encouraging the industry, so that the industry can generate innovative health products and make them available to whoever needs them (WHO, 2011b).

The regulation phase consists in protecting the society by publishing rules, rehearsing protocols, pre-authorizing purchases, registering, post-sale vigilance and reporting on contra-indications. The focus in this phase is on guaranteeing the safety of patients and technology users (WHO, 2011b).

In the HTA phase, it is possible to systematically evaluate the proprieties, effects and/or impacts of ME on the deliverance of health care by a well-designed and defined methodology. The main target of this phase is to educate the health policy formulators on related technology. Thus, it is possible to properly plan from incorporation of the ME to the removal of ME. Depending on the issues involved, time frame for decision making and availability of resources, the HTA can be tackled in different ways, such as by more detailed HTA reports, availing of reports produced elsewhere, fast review, and monitoring technological reports (Velasco-Garrido & Busse, 2005; HTA GLOSSARY, 2010).

The HTM phase encompasses a variety of attributions, which include, *inter alia*, the following: identification of needs, collection of reliable data about ME, incorporation process, a complete inventory of ME, maintenance program based on risk reduction and safe operation, aiming for safe tools and high quality health service, allotting sufficient resources to maintain the technology under use, monitoring the clinical effectiveness of ME, updating and deactivation or replacement of unsafe and obsolete equipment (Kaur, 2005b; Santos & Garcia, 2010; WHO, 2011b).

From the foregoing discussion, it follows that each phase has specific attributions. However, it is important to highlight that technology life cycle phases are not independent; that is, action taken in one phase may impact other phases. This underlines the need for adequate planning of ME incorporation, and thus for strengthening technology life cycle, the actors involved and the health system.

The technology life cycle phases can operate at local, regional and national levels. The characteristics, perspectives and impacts of each phase are described in Table 1.

	R&D	Regulations	HTA	HTM
Perspective	Innovative knowledge, application and tools for health services	Safety & efficacy	Population served	Health services provider
Orientation	Personal health services	Population safety	Population health	Community health services
Requirement (Output)	Improved and/or new tools & services	Mandatory compliance	Recommendations on highly complex technologies	Operational rules and guidance for all medical devices
Method	Innovation and improvement	Performance testing, safety assessment & post-market reporting	Systematic analysis, critical review	Operational management of technology life-cycle
Criteria	Market adoption	Safety and quality standards	Epidemiology data, statistics, analysis of efficacy, effectiveness, and appropriateness	Needs analysis, specifications, reliable device availability for clinical use
Outcome	Enhanced health services	Risk mitigation and prevention of harm	Responsiveness and maximization of clinical outcomes and cost-effectiveness	Improved health delivery; sustainable availability of high-quality and safe devices

Table 1. Characteristics of ME life cycle phases (Source: WHO, 2011b).

2.1.3 Actors involved in the incorporation process

Past experience shows that ME incorporation process has not been well coordinated in most countries. In many cases, it led to undesirable results, such as increase in cost, abusive use of the facilities and frustrating managers, users and patients. However, by learning from these experiences, the application of knowledge has been so conditioned as to derive maximum advantage from each of the ME life cycle phases (David & Judd, 1993; Sprague, 1988).

The incorporation process, which involves clinical, technical, financial, infrastructural and human resource impacts, is a challenging task. Therefore, a multidisciplinary team is required to plan and execute the ME incorporation process effectively. The team can prevent recurrence of past errors, and identify the potential factors that may lead to dissatisfactory results. The team must be formed by including representatives from clinical, administrative, financial, clinical engineering, installations, information technology and material management areas. Besides, it needs to be strengthened with specialized knowledge and services of consultants and distributors (Coe & Banta, 1992; Wang, 2009).

The Clinical Engineers can be strong members of the team in that they ensure that the real clinical needs of the user are identified, treated and, when possible, attended to. They can serve as a communication link between professionals of different disciplines involved in the incorporation process, inside or outside the HCC. Besides, their experience and skills can be used to help the HCC in its systematic and safe incorporation of the technology process. It is

fundamental that those Engineers never forget that they are just members of a team and not the only ones responsible for ME incorporation (Harding & Epstein, 2004).

User's integration in the development and assessment of ME is explicitly recommended in literature. This perspective turns out to be beneficial to technology producers, besides highlighting the importance of users inside the incorporation process (Woodside et al, 1998; Kittel et al, 2002; Sarwar & Robinson, 2007). In addition, other approaches directly reflect on the potential impact of the user's integration into the assessment process. (Mcgregor & Brophy, 2005).

Just as Clinical Engineers and users, all other actors in the incorporation process—interns, like patient groups, or externs like manufacturers, distributors or regulators—need to be considered equally important, and treated accordingly(Gibson et al., 2004).

Regarding decision making in ME incorporation, the representation of multiple perspectives of the actors is a key element of justice (Singer et al., 2000). Similar approaches can be found in Drummond et al. (2008), which presents the key principles to guide the HTA.

2.2 Identification of resources for incorporation

It is important to stress that, without necessary resources, adequate planning of ME is not possible, and consequently the incorporation process too. The progress of health technology with assured benefits to the patients and increased efficacy is rather slower within the health systems than in other health service economies. This can be ascribed to several barriers in this process, such as the following (Robert et al., 2009):

- Lack of formal mechanisms to disseminate recommendations and information about ME assessment.
- Availability of adequate data on the cost and price of new health technology.
- Insufficient sharing of information between buyers and sellers that can result in bad purchasing decisions.
- The culture within the health systems is not sufficiently entrepreneurial.
- Lack of financial and technical support to the companies in turning innovative ideas into marketable products.
- Bureaucracy around purchasing procedures.
- Need for training the health system teams in using the new ME.

Fortunately, notwithstanding these barriers, the new ventures introduced by World Health Organization, along with recommendations towards ME, have been well accepted by the country members. These can be turned into better and more efficient health systems. Nowadays, in developing countries like Brazil, some attention is being given to preparation and dissemination of methodology guidelines for assessment and incorporation of ME. This contributes to the development of recommendations to deal with the challenge of ME incorporation process.

Even so, many barriers still remain to be broken down to achieve complete success in ME incorporation and usage. In this context, it is important to identify clearly the needs for adequate planning. One of the most critical necessities is undeniably the human resources. In general, the two common mistakes are excessive centralization of decision-making and

reposing too much confidence in the specialists concerned. Centralizing decisions brings political problems related to favoritism, subjectivity, and lack of transparency. And, too much confidence in specialists needs credibility and general support, which sometimes get worse because of lack of a wider view (Wang, 2009).

Therefore, it is necessary to establish a transparent and efficient process that can identify and plan the actions pertinent to ME incorporation process. As has already been said, the engagement of many actors in this challenge can bring more transparency and a generalized approach, as well as all the relevant factors. So, it is recommended that a multidisciplinary team be formed and supported by representatives of every group of actors involved, where the Clinical Engineer can act as a task facilitator.

However, it is necessary to get information and evidence of internal and external sources of ME, which can enable the planning of ME incorporation. Following are some internal sources of information and evidence and the main factors to be considered for each source (Wang, 2009):

- Current users: efficacy, effectiveness, safety, easy training and usage.
- Clinical Engineering: reliability, safety, maintenance and availability.
- Installation management: requirements of usefulness and environment impact.
- Information Technology: network problems and software support.
- Material management: input, accessories and alternative distributors.

Following are the external sources of information and evidence, and the main factors to be considered (Wang, 2009):

- Health Information Centers: epidemiological data, possible refund, rules and regulations, marketing rivalry, financial problems and HTA reports.
- Manufacturers: product specifications, financial conditions, requisites to installation and functioning, post-purchasing guarantee and support.
- Regulating mediums, Civil Engineers and Architects: infrastructure requisites and impacts, regulations and codes.
- Other distributors: aiding equipment and furniture, alternative supply and service sources.

The multidisciplinary teams at the local level can be considered as determined people to ME incorporation management in the health system as a whole. However, their attributions and actions must be tailored to the needs of the situation. The actions of the multidisciplinary team at local level need administrative and technical support. Systematic research of scientific literature and HTA reports by specialized technicians would be helpful for simultaneous execution of the planning tasks within the timeframe given for ME incorporation process (Kaur, 2005a; WHO, 2011a).

2.3 Evaluation of the necessity for medical equipment incorporation

Prior to ME incorporation, one must clearly understand the difference between desire and need. This is because many acquisitions were made more under the impulse of desire and for subjective reasons, than in the common interest of the majority of the actors, which must be the case. The need for ME incorporation must be assessed rationally by discussions with

the clinical team on the diseases that need to be addressed and the health policies they recommend, and not the technology they want (Wang, 2009).

Besides, a survey must be undertaken to check if the HCC that is going to receive the technology has already some ME that meets the clinical needs under consideration, or if it has any unused equipment. Following are the other questions that must be taken into account in this regard (Kaur, 2005a; WHO, 2011a; Robert et al., 2009; Wang, 2009):

- Does the demographic density of the HCC region that is going to use the ME justify the incorporation?
- Does the ME have an entry in the register of competent regulating establishment?
- Is there any demand in the health service offered by the ME?
- Is there any personal preference from the clinical or administrative team of the HCC?
- Will the ME incorporation and its results impact significantly on the treatment/diagnosis of patients by any other specialist?
- Does the complexity of the identified health problem justify ME incorporation?
- Is there any guarantee that the manufacturer will offer spare parts during the projected life cycle of the ME?

All these issues justify a detailed ME assessment, which needs time, human resources and financial investments. If the majority of the factors justify the need for ME incorporation, one can pass on to the next planning phase, which involves assessment of the impacts upon users, patients, infrastructure and immanent traits of technology. In case the need is not justified, immediate disengagement of the incorporation process must be considered.

Answers to the proposed questions can be found in international and national literature. Yet, the technological park of HCC may have to be covered to check if any unused ME exists. Data in respect of demographic density and demand can be taken from the Health Ministry sites. Interviews with clinical and administrative team, as well as with some manufacturers, are recommended to identify personal preferences and ascertain the capacity to supply the spares, respectively.

When applicable, it might be necessary to assess the amount of ME needed, on the basis of epidemiological data, population to be assisted, geographic distances to be covered, status of the HCC that is recently in need of technology, including its capacity to utilize the equipment or the usage time per case. However, sometimes, it may not be possible to go beyond just the figures, because many variables are subjective or difficult to estimate. Besides, the available data might not be reliable enough to lead to correct indications. Nonetheless, some attempts can be made, assuming potential risks and making adjustments, so that they provide some basis for future assessments (Wang, 2009).

2.4 Impact of medical equipment incorporation

The impact of technology incorporation into health services, particularly of ME, can be viewed in both positive and negative ways. The determining point of the impact will be the way in which the planning is conducted. Therefore, before incorporating ME into the health systems, one must study the likely impact of this equipment on the service, both direct and indirect. One of the main reasons cited for the disuse of ME is sometimes the failure to predict ME's impact (World Health Assembly Health Technologies [WHA], 2007). These

impacts can be portrayed as three pillars: Human Resources, Technology and Infrastructure. Figure 1 shows the impact on health technology from Clinical Engineering view, particularly ME:

Fig. 1. Impact on health technology, particularly ME, from clinical engineering view must be considered to obtain safety, reliability and efficiency deliverance of health service by ME usage (Source: Santos, Souza & Garcia, 2010).

2.4.1 Impact on human resources

Health professionals are individually responsible for the transparency of their practices in certain aspects of health care offers. Therefore, they have the responsibility, as part of their continuous professional development, to acquire, maintain and disseminate knowledge and abilities in availing of ME. Before inducting health technology into HCC, the managers must ensure that the health professionals are adequately educated to guarantee safe usage of technology (NHS, 2005).

The users' training needs can cover educational services, as well as clinical users' training. Safety training aspects, such as those with laser equipment, must also be considered for inclusion in user training needs (Harding & Epstein, 2004).

Additionally, a training plan is necessary, considering the training material, manuals, trainers and other resources pertinent to the training, as also the need of the establishment of a schedule of personal training activities in order to regard the personal turnover and gradual loss of competence. This plan must take into account some fundamental aspects (NHS, 2005) listed below:

- Degree of ME risk and, therefore, priority level.
- The need for flexible approaches to learning.
- Accessibility to all ME users.
- Constant information about the changes made in the legislation pertinent to ME.

However, for conducting any program one incurs cost, which can be directly related to the learning curve of the user in relation to the ME which will be used.

The learning curve is a tool which can monitor the performance of workers assigned with certain tasks. Through the curves, it is possible to evaluate and plan for more productive tasks, and thereby, to reduce the loss arising out of the inability, which is checked, above all,

in the first periods of implementation (Dar-El, 2000). The tool also allows adequate allocation of tasks to the members of the workgroups so as to enable them complete their performance characteristics, besides the monitoring of costs related to the process (Anzanello & Fogliatto, 2007).

Figure 2 shows the relation of cost versus ME complexity, divided into two categories (A & B). The line 0b represents the cost or time spent to train a technician (beginner) in operating the equipment of category B and the line ba in operating the equipment of category A. From their comparison, it can be seen that the more complex the ME is, the more would be the time (or cost) required to train the professional (Souza et al, 2010; Cheng, 2006).

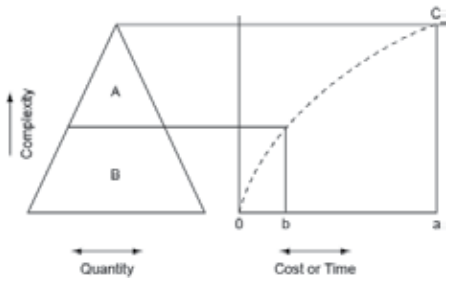


Fig. 2. Training curve based on ME complexity (Source: Cheng, 2004).

In terms of the magnitude of complexity, one can consider classifying the ME based on the following definitions (Calil & Teixeira, 1998):

- Low complexity equipment: The ME of this category has complex electronic or mechanical circuits, but they pose no maintenance problem (e.g., thermal double boiler, sterilizer, sphygmomanometer, mechanical scales, etc.) Those who operate this equipment need not be specialists, and the training they need is quite simple.
- Medical equipment of medium complexity: The ME of this category requires personnel with basic education and training that can meet the repairing needs. Examples of the ME of this category are incubator, centrifuge, cardiac monitor, electrocardiograph, hemodialysis equipment, etc.
- High complexity equipment: The ME of this category demands qualified technicians with specialized training. In many cases, these technicians have higher education and some of them had foreign training. Following are some examples of this equipment: nuclear magnetic resonance, scanner, chemical analyzers (some types), gamma chamber, linear accelerator, ultrasound machine (image diagnosis system), etc.

Therefore, the degree of complexity of ME can help in estimating the costs and the training time required for each ME, because the more complex the ME is, the higher would be the cost and time needed. In other words, from the degree of complexity of the equipment, one can draw a qualitative estimate of the cost and time required to train a person in operating that equipment. For instance, the cost and time required to manage an ultrasound machine, which belongs to the high complexity category, would be much higher than the cost and time required to operate a cardiac monitor that belongs to the medium complexity category.

With that information in mind, it is possible to properly hire specialized training services or even sign maintenance contracts for users' training. However, final cost estimates can be

made only after a market survey and discussions with manufacturers, distributors, and companies specialized in ME training.

2.4.2 Impact on technology

The ME needs to be operated in an efficient and safe way. To achieve this, various factors that may interfere with each other will have to be considered. So, one must prepare a maintenance plan that covers not only preventive or remedial maintenance, but also detects potential and hidden errors that are not usually identified by users, but can cause injury or death to the patients (Kaur, 2005a; Wang, 2009).

For preparing a maintenance plan, one must consider the following actions (Kaur, 2005a):

- Check the guarantee date and enquire if the distributor offers, during the guarantee period, the spares required, and if the guarantee period can be extended for an acceptable cost.
- Check, in case of any breakage of ME, whether the manufacturer will replace or repair the broken part or even offer refund if the equipment has manufacturing or material defects. Will the offer cover all parts of the equipment? Does the manufacturer pay for the shipping expenses?
- Ensure availability of consumables, accessories, spare parts and maintenance materials.
- Check if the maintenance requires the service of a qualified engineer, and if the answer is 'yes', identify the local distributor or representative who can help in case of breakdown or glitch.
- If no distributor or representative is available locally, check if somebody is available at regional or national level.
- To increase the bargaining power for entering into a maintenance contract, check if there are companies, other than the authorized agent, who can offer maintenance service.
- Identify, from the options available, the maintenance contract that has the best cost-benefit ratio for each ME. In most cases, purchasing ME by lending or leasing is advantageous, but it needs to be checked if the input and maintenance costs do not exceed the purchasing costs in a short time.

It is important to note that the ME, which does not have adequate support of maintenance services, consumable goods, and replacement parts, it is probable that the ME may remain unused for long periods and might ultimately be replaced prematurely. Therefore, it is essential to any health establishment, no matter its size, to implement a ME maintenance program. The complexity of this program depends on the size and type of installation, its locale, and necessary resources. The need for a good maintenance program will be the same regardless of whether the ME is in a high income, urban environment or in a low or medium income, rural environment (Kaur, 2005a; WHO, 2011c).

The ME maintenance can be divided into two categories: Inspection and Preventive Maintenance (IPM), and Corrective Maintenance (CM) (see Figure 3). IPM includes all programmed activities that guarantee equipment functionality and prevention of failure[3].

[3]The condition of not meeting intended performance or safety requirements, and/or a breach of physical integrity. A failure is corrected by repair and/or calibration (WHO, 2011b).

Inspections of performance and safety verify the functionality and safe usage of a tool. Preventive Maintenance (PM) refers to the programmed activities to ensure that the ME endures its useful life through actions like calibrating, replacing dysfunctional parts, greasing, cleaning, etc. Under PM, the inspection can be done as an individual or group activity to guarantee ME's functionality. CM refers to activities carried out to restore the physical integrity, safety and/or performance of a failure ME (WHO, 2011c).

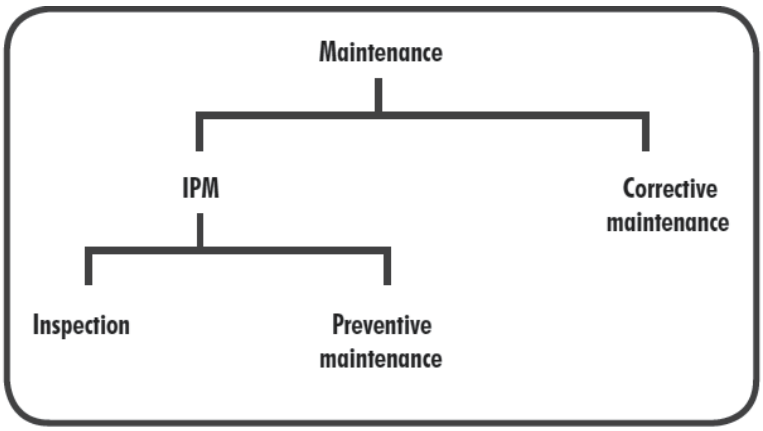


Fig. 3. Categories and types of ME maintenance: Inspection and Preventive Maintenance; and Corrective Maintenance (Source: WHO, 2011c).

2.4.3 Impact on infrastructure

Once the ME is incorporated into HCC, it is important to understand different aspects of the resulting impact on the infrastructure. Following are some of the advance actions to be carried out before acquiring the technology (Calil, 2007; Wang, 2009; WHO, 2011d):

- The space needed to install the ME.
- The type of floor, and equipment weight and disposition in relation with other technology equipment in adjacent rooms.
- Type, size, and position of the place and building.
- Check if the HCC has more than one floor, and if 'yes', identify the floor for installing the ME.
- Ascertain the availability of gas and water supply and their supply conditions, like type, quality and quantity, and pressure.
- Check the availability of power supply for electric connections; also, check if the HCC has an emergency generator.
- Check the need for weatherproofing such as air-conditioning, and quality control (air quality and humidity).
- Other factors that may be specified as installation prerequisites.

These actions are particularly important to HCCs in rural areas and developing countries where stable sources of energy, adequate water supply and controlled environment in terms of temperature and humidity are not always available.

Information about the likely impact on the infrastructure can be obtained from the manufacturers. They usually offer architectonic projects and support structuring layouts for installing robust ME, such as robotic surgery system and magnetic resonance instrument (Wang, 2009).

2.5 Proposed model for medical equipment incorporation

Clinical Engineering plays an important role, through Health Technology Management (HTM), in innovation, incorporation, usage/utilization and ME re-processing. Thus, the proposed model comes from HTM incorporation phase. It is important to highlight here that, in the last few years, the profile has been undergoing some changes in the incorporation process, which are not being released just in the HCC, but also in the entire health system. Therefore, the methodology aimed at helping this process must contemplate taking such actions that can be applied to the benefit of public health (Sônego, 2007; Santos & Garcia, 2010). Figure 4 depicts the conceptualization of the proposed model, with focus on ME incorporation phase.

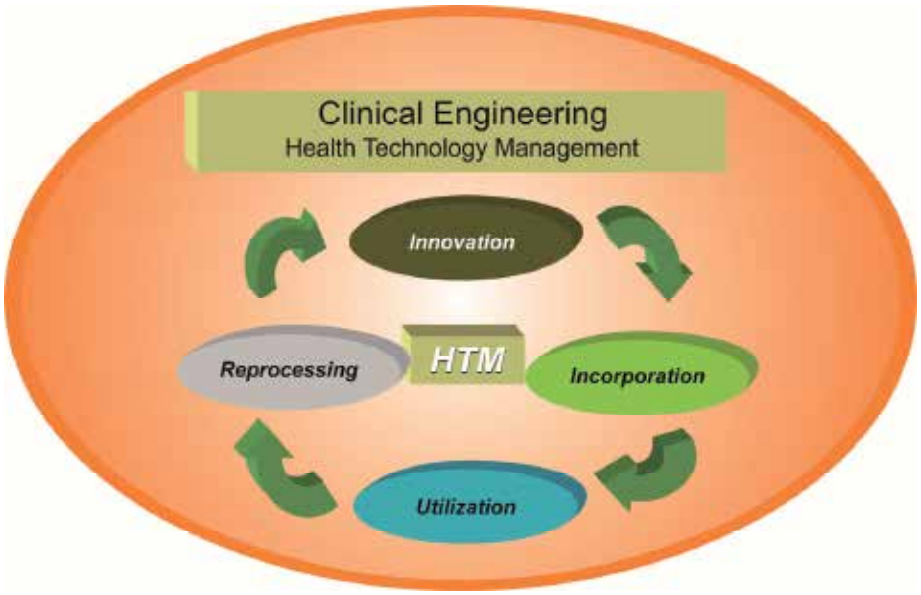


Fig. 4. Conceptualization of the proposed model with focus on ME incorporation phase.

It is important to note that the phases of ME life cycle are not independent, i. e., the actions in any one phase can impact the other phases. Besides, each phase has specific stages, which must be appropriately planned and guided to obtain satisfactory results in respect of the patients. Thus, by monitoring the actions carried out in one phase of technology, and observing the consequent impact on other phases, one can plan in a better way the actions in other phases of the ME life cycle.

Against this background, a model was developed to support the ME incorporation process, as shown in Figure 5.

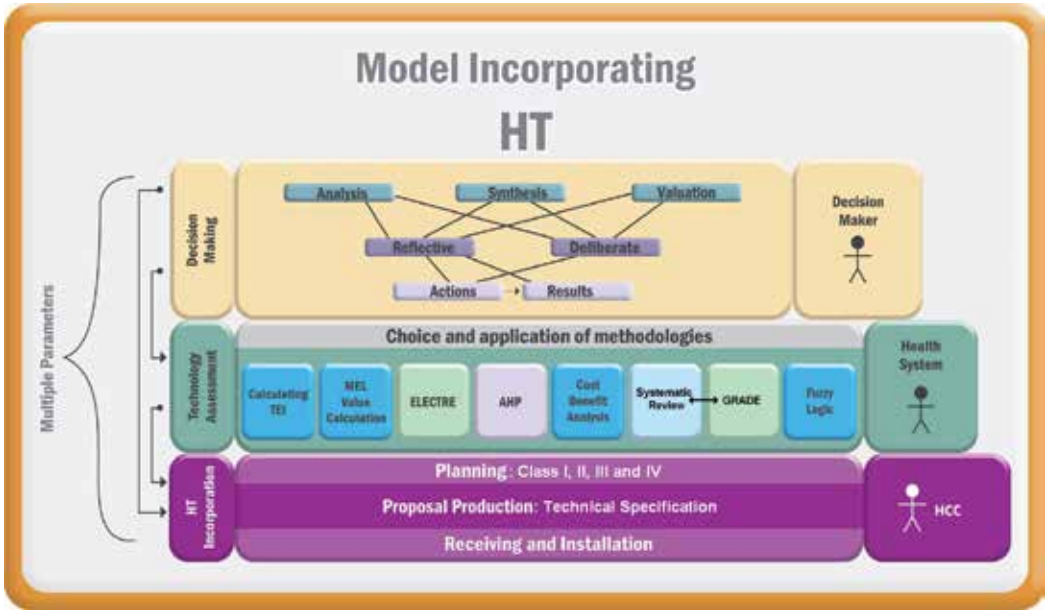


Fig. 5. Model depicting the process of health technology incorporation process, particularly ME. The proposal is based on decision making domains, technology assessment and HT incorporation. The domains establish interconnections, and there are multiple parameters and actors to be considered.

In the decision making domain, interconnections are made among the mind functions, decision making types and actions. The mind functions are made by analysis, synthesis and imagination, and evaluation. The analysis consists in separating a whole into its constituent parts, the synthesis and imagination is the reverse of analysis, that is, it presents or puts the things in groups to make a whole. Evaluation comes into action in mental activities, such as success criteria establishment, performance evaluation and judging people (Adair, 2007).

These functions can relate with those decision making types that can assume a reflective or deliberated form. The reflective one comes from the necessity of reflecting on how people make decisions based on their experiences, where they use the knowledge acquired by experience to identify and evaluate the situations and later make decisions (Fadok, Boyd & Warden, 1995; Zsambock, 1997). The deliberated form is based on reason, which supports the decision process, wherein the decision maker within his/her context will analyze, synthesize and evaluate to achieve the desired result (Klein, 1997; Joseph, 2007). So, from the interconnections between the mind functions and the decision making form, actions and post results can be created with the decision maker as the actor.

The decision analysis, which sets relevant technological alternatives, together with systematic review of studies about the effects of technology on health management, and the economic analysis that relates costs and effects, forms the main methodology used in HTA (Krauss-Silva, 2004).

Within the technology assessment domain, it is possible to choose and apply multi-criteria methodology to support decision making, and assessment methods of clinical evidences and

costs. The methodology that can be employed in the ME assessment includes, *inter alia,* the following approaches: calculation of Maintenance Expended Limits (MEL) value; economic analysis; Elimination and Choice Translation Reality (ELECTRE); Analytic Hierarchy Process (AHP); Multi-Attribute Failure Mode Analysis (MAFMA); Measuring Attractiveness by a Category Based Evaluation Technique (MACBETH); fuzzy logic; systematic review; Grading of Recommendations Assessment, Development and Evaluation (GRADE).

Lastly, the HT incorporation domain includes three stages: planning; proposal production; receiving and installation. These stages show multiple parameters to be evaluated. In the planning stage, four classes of parameters must be considered. These are Class I: Safety; Class II: Efficacy/Effectiveness; Class III: Infrastructure Impacts, Human Resources, Maintenance and Regulatory Aspects; and Class IV: TCO and Economic Analysis. The proposed production stage must enable technical specification of the technology to meet the clinical and technical needs. At the time of receiving the technology, one must check if the technology satisfies the technical specifications, and if it has all the essential accessories. Only after that, the equipment can be installed as planned and commissioned after performance and safety tests.

Thus, the architecture of the model depends on the inter-relations the domains, where the HT incorporation domain relates to the decision making one and the technology assessment one has multiple parameters and actors involved in the process (decision maker, health system and HCC).

2.5.1 Surveying parameters to be evaluated

Initially, the researcher must verify whether the foundation available for ME incorporation is strong enough to justify detailed assessment of health technology. This is because a thorough assessment of ME needs time, specialist professionals and consequently investments. If the foundation is found unfavorable, the assessment can be aborted with justifications based on the same questions raised for eligibility assessment, following the issues against item 2.3. However, depending on the context and technology under assessment, the incorporation team might ask additional questions during eligibility assessment for an initial map of the technology and thereby avoid wastage of time and investment over unjustified assessment or even unnecessary ME incorporation.

If the eligibility is found to be favorable to technology, one must identify the phase in which the ME is. This is important, because the ME in the adoption or incorporation phase will possibly have higher variation in clinical effect in comparison with that of the ME in usage phase. The ME in the obsolescence phase must be avoided, because of non-availability of spares and high maintenance costs.

It is important that this survey considers the life cycle phases of technology right from the acquisition phase. The ME identified as belonging to previous phases, such as 'under development' and 'pre-market', which are relevant to the health system, must be evaluated by the *Early awareness and alert systems* of EuroScan (Simpson et al., 2009). This is because the technology under the previous phases of incorporation has specific characteristics, mainly in relation to scientific safety evidence and efficacy/effectiveness, as the ME effect has not been observed in large scale.

After assessing the eligibility and identifying the life cycle phase of the ME under consideration for incorporation, one must undertake a more detailed assessment of the technology. This assessment must satisfy all aspects of Classes I, II, III, IV considered in the planning phase of incorporation, as shown in the model presented in Figure 5.

Classes I and II are essential in the evaluation process, because there must be evidences of safety and efficacy/effectiveness that satisfy at least the minimum conditions for using the technology without causing any harm to the patients and users. These parameters must be evaluated from the clinical data available in literature, systematic reviews or HTA approaches, such as the following: "*Methodological Guidelines: Health Technology Assessment Appraisals*" "(BRASIL, 2009a), "*Clinical Evidence for Medical Devices: Regulatory Processes Focusing on Europe and the United States of America* "(WHO, 2010) or "*Health Technology Assessment Handbook*"(Jørgensen, 2007; Stenbæk &Jensen, 2007).

In a rational sense, the ME incorporation team must satisfy itself about the quality of the available evidence and assess whether the criteria of safety and efficacy/effectiveness meet the minimum acceptable conditions required to proceed with the assessment of the ME applicant for incorporation. Failing this, it makes no sense to evaluate other aspects.

However, if the results obtained in Classes I and II are favorable, one must try evaluating Classes III and IV. Class III covers different aspects of infrastructure, human resources, maintenance and regulatory procedures. The investigators are encouraged to consider four essential factors in this regard:

- Learning curve: This criterion refers to the time and effort required to train a user in effective use of the ME. For estimating these, the complexity of the ME must first be assessed, because the more complex the ME is, the more would be the time and effort required to train operators and technical team. For information relevant to this criterion, one must check with the distributors, manufacturers, similar ME inventories, and establishments that publish technical manuals about ME, such as ECRI.
- Installation ease: Installation ease is linked to infrastructure conditions, which may include alteration of physical space, adaptors, accessories, compatibility with other technologies, energy, water and gas supply nets, humidity and temperature controls, and input storage needs. Information relevant to these aspects can be obtained from regulatory establishments and sometimes manufacturer's manuals.
- Maintenance ease: This criterion covers all the conditions necessary for executing the maintenance plan. Foremost among them is the availability of professionals in the region, state or country, who can train the technicians and technology users in operating and maintaining the ME to be installed. Besides, one must also ensure a suitably worded guarantee, availability of spares, facilities for software updating, indigenous availability of authorized distributors, and possibility of finding a third party for maintenance through a contract that is linked to the purchasing of goods or even the renting of ME. The most important thing in meeting these requirements is to identify the best cost-benefit ratio that calls for no compromise in meeting the clinical needs, and the one that ensures optimum utilization of the useful life time of the ME.
- Usability: As far as this criterion is concerned, no single technique can answer all the questions. Therefore, what is needed is a combination of techniques, considering the

medical environment limitations, and the human costs in terms of fatigue, stress, frustration, discomfort and satisfaction, learning talent, ME use tax, adaptability to the task and the user's needs, and user's characteristics. To achieve global usability, one must address the following measures:

- Efficacy: Percentage of aims realized, and of users who completed the task successfully, and the average of completed tasks.
- Efficiency: Time to complete a task, tasks completed per unit of time and monetary cost of task realization.
- Satisfaction: Satisfaction scale and frequency of use and complaints.

The usage measures cited above (or their estimates) can be obtained by interviewing the clinical team or from similar ME inventories, and pre-market study reports submitted to the departments concerned by the registrar of commerce, for example, ANVISA[4] and FDA[5].

As regards Class IV costs, the criteria that must be considered are those, which might be covered by the TCO and economic analysis. The TCO corresponds to the sum of the costs of acquisition, operation, maintenance, training and replacement. Calculating these costs is sometimes challenging. Therefore, they might be estimated on the basis of information gathered from the distributors, manufacturers and HTA reports. The idea behind estimating the total property cost is to ensure that one does not go just by the acquisition cost, which can be attractive, but also other costs that might go against technology usage.

Through economic analysis, one can investigate the cost-benefit relation to ascertain if the results obtained from the technology under assessment justify the costs, and whether they compare favorably with other technological options that show good cost-benefit relations. Instructions on economic analysis with focus on health technologies can be obtained from the guide "*Methodological Guidelines: Economic Evaluation of Health Technologies*" (BRASIL, 2009b).

These guidelines would be helpful in undertaking the team activities of ME incorporation. However, with the help of the Clinical Engineer, one can add another criterion. The incorporation process will be a challenging one in the context of variations related to geographical regions, health policies, demand, human and financial resources, and cultural aspects, among other pertinent factors.

3. Conclusions

One of the factors that reaffirms the importance of ME incorporation is the improvement in people's health during the last decade, an achievement that could not reach the poor and other socially marginalized or excluded groups earlier. Increasing inequalities in health status are more evident in rural areas. This situation was created by the uneven distribution of money, power and other resources at global, national and local levels, which were in turn influenced by political equations. The health social determinant is mainly responsible for the inequalities in health. The available evidence points to a two-way relationship between poverty and health. Within this vicious circle, poverty creates poor health, and poor health

[4] http://portal.anvisa.gov.br
[5] http://www.fda.gov/

creates poverty. Within the vicious circle of higher income is good health and good health is related with higher income and welfare (WHO, 2011b).

The Clinical Engineering can contribute, through administration and management of ME, to the preparation, guidance and observation of the impact of methodologies aimed at ME incorporation planning in the HCC in the context of health systems. Additionally, adequate incorporation planning requires multidisciplinary knowledge; so, the Clinical Engineer, who has multidisciplinary education, can act as a facilitator by establishing an interface among the actors involved in ME incorporation and by promoting the culture of constant monitoring of the impact of technology on health, after its incorporation in the health system. The observation of the impacts and the lessons learnt from past and recent needs can contribute to planning future incorporation, (Moraes, 2007; Santos & Garcia, 2010; Signori & Garcia, 2010).

Many times, technology management in health is seen as an independent task, but for a few links with other parts of the health service. In other words, in the past, the technical personnel were hardly ever involved in crucial activities such as investment plans, service quality evaluation or organizational issues. However, this scenario has been undergoing some change in the last few years. So, ME management can now be clearly defined as an integral part of the health system and its activity felt at all levels of the public health service (KAUR, 2005a).

The ME cannot be managed in isolation, but only with other components of the health care, including the aims, procedures, finances, level of personnel and support systems at each health service level. To accomplish this, the creation of a multidisciplinary group of management of technology is recommended for each level (local, estate, and national). This group must have representatives of different disciplines: medical, clinical, clinical engineering, support service, purchasing sector, financial and maintenance team of ME (KAUR, 2005a).

Within this context, the incorporated ME is fundamental to health care service, particularly to diagnosis and disease treatment. The available and accessible ME in health care environment is related to the equity and health service offer that is more relevant to the patients' needs. Any national health plan needs policies, strategies and plans of action to health technologies, especially the ME. A robust health system must guarantee access to safe, efficient and high quality ME, in order to prevent, diagnose and treat diseases and injuries, and help the patients in their rehabilitation, and to promote public health (WHO, 2011b).

ME incorporation is an important element of the HTM. This is a complex and multidisciplinary process for developing the activities to support decision making, though some members of the health team and distributors believe that it is just the action of purchasing. For example, costs outside the budget for additional accessories may become necessary after the supply order for ME has already been placed. Or, some unexpected changes may become necessary in installation plans, because the dimensions and other specifications of the ME have not been properly worked out. This entails considerable costs and delays, besides impairing the quality of the public health system. Yet, the technology may remain completely unused, consequently its use can harm the patient or personnel, thus impacting the public health in a negative way (Harding & Epstein, 2006). It underlines

the need to systematize the ME incorporation process and thereby mitigate or eliminate some negative factors of the process that can affect the technology life cycle.

The model proposed here is based on constructing three domains: Decision Making; Technology Assessment and ME incorporation. It can help the incorporation team in identifying, predicting and guiding the realization of measures to minimize possible unfavorable impacts, as also to maximize the benefits obtained through ME incorporation. This is possible because the model has been built on scientific evidence and reliable information available in literature. Besides, the proposed model would be helpful to future researches in that it represents a consolidated methodology that deals with the multiple parameters involved in the ME incorporation process in a systematic and rational way. The Clinical Engineer, as a multidisciplinary education professional, can be a fundamental actor in methodology development and a facilitator of ME incorporation, which can ensure health service deliverance in a safe, effective, and equitable way, besides rational utilization of the resources in the developing countries.

One can observe that health technology, particularly ME, suffers from lack of clinical evidences in the innovation and incorporation phases of the life cycle. Besides, a higher variation is expected on the clinical effect as compared with the technology under wide usage. This is because the technology that belongs to the initial phases has not been monitored on a large enough scale. Therefore, at the time of incorporation, one must prioritize the ME that is in the usage phase. The Clinical Engineering can be helpful to the incorporation team in the identification phase of the life cycle, as well as in the search, selection and assessment of available evidences in literature so as to ensure that the technology to be incorporated is safe and effective. So, the deliverance of health system with quality and equality contributes to the promotion of public health (Sônego, 2007).

Additionally, it is important to highlight that the phases of technology life cycle are not independent, i. e., actions in any one phase can impact other phases. For example, an inadequate incorporation plan can lead to high costs of maintenance, unavailability of technology, risks to patients and users, and spilling of unsatisfactory clinical results into other phases of the ME life cycle (Moraes, 2007).

Lastly, one must note that, after ME incorporation, the Clinical Engineer retains the management of other phases of technology life cycle with him for future ME updating, improvement and replacement. The services of Clinical Engineering are more necessary at this stage to deal with the responsibility of ME management program within the framework of guidelines that range from the strategic phase to the replacement phase.

4. Acknowledgment

The authors are thankful to the CAPES (Coordination of Improvement of Higher Education Personnel) for financial support, and IEB-UFSC, for motivation and support to this research.

5. References

Adair, J. E. (2007). *Decision Making & Problem Solving Strategies* (2nd ed), Kogan Page, ISBN 10 0 7494 4918 7, Philadelphia, USA

Anzanello, M. J., & Fogliatto, F. S. (2007). Curvas de Aprendizado: Estado da Arte e Perspectivas de Pesquisa. *Gestão da Produção*, Vol. 14, No. 1, (jan.-abr. 2007), pp. (109-123)

BRASIL. Ministério da Saúde. (2009a). Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Ciência e Tecnologia. Diretrizes Metodológicas: Elaboração de Pareceres Técnico-Científico, In: *Instituto Nacional do Cancer (INCA)*, 03 September 2011, Available from: <http://www1.inca.gov.br/inca/Arquivos/publicacoes/diretrizes_PTC_2_edicao_2009.pdf >

BRASIL. Ministério da Saúde. (2009b). Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Ciência e Tecnologia. Diretrizes Metodológicas: Estudos de Avaliação Econômica de Tecnologia em Saúde, In: *Biblioteca Virtual em Saúde. Ministério da Saúde*, 03 September 2011, Available from: < http://bvsms.saude.gov.br/bvs/publicacoes/avaliacao_economica_tecnologias_saude_2009.pdf >

Calil, S. J. (2007). Caminhos para a Incorporação de Tecnologias em Saúde. *Debates GV Saúde*, Vol. 3, (Primeiro Semestre de 2007), pp. (31-34)

Calil, Saide J.; Teixeira, Marilda S. (1998). Gerenciamento de Manutenção de Equipamentos Hospitalares. Série Saúde & Cidadania, v.11, pp. (8-47).

Coe, G. & Banta, D. (1992). Health Care Technology Transfer in Latin America and the Caribbean, *International Journal Technology Assessment Health Care*, Vol. 8, No. 2, (March 1992), pp. (255-267)

Cheng, M. (2004). A Strategy to Maintain Essential Medical Equipment in Developing Countries, In: *Clinical Engineering Handbook,* Joseph Dyro, pp. (133-134), Academic Press, Retrieved from<http://www.elsevier.com/wps/find/bookdescription.cws_home/702695/description#description>

Centers for Medicare and Medicaid Services. (2000). Review of Assumptions and Methods of the Medicare Trustees Financial Projections, In: *Technical Review Panel on the Medicare Trustees Reports*, March 2011, Available from: <https://www.cms.gov/reportstrustfunds/downloads/TechnicalPanelReport2000.pdf>

Cutler, D. M., & Mcclellan, M. Is Technological Change in Medicine Worth it?. *Health Affairs*, Vol. 20, No. 5, (September 2001), pp. (11-29)

David, Y. & Judd, T.M. Medical Technology Management, Space Labs Medical (Medical Inc. Biophysical Messurement Series), Redmond, WA, 1993

David, Y., & Jahnke, E. G. (2005). Medical Technology Management: From Planning to Application. *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference,* Shanghai, China, September 1-4, 2005

Dar-El, E. (2000). *Human Learning: from Learning Curves to Learning Organizations* (1 edition), Springer, ISBN-10: 0792379438, New York

Fadok, D. S., Boyd, J., & Warden, J. (1995). Air Power's Quest for Strategic Paralysis, In: *School of Advanced Air Power Studies*, 12 April 2011, Available from: <http://dodreports.com/pdf/ada291621.pdf>

Gibson, J. L., Martin, D. K., & Singer, P. A. (2004). Setting Priorities in Health Care Organizations: Criteria, Processes and Parameters of Success. *BMC Health Services Research*, Vol. 4, No. 1, pp. (17-25)

Harding, G., & Epstein, A. (2004). Technology Procurement, In: *Clinical Engineering Handbook,* Joseph Dyro, pp. (118-122), Academic Press, Retrieved from < http://www.elsevier.com/wps/find/bookdescription.cws_home/702695/description#description>

HTA GLOSSARY. International Network of Agencies for Health Technology Assessment and Health Technology Assessment international, August 2010, Available from <http://www.htaglossary.net>

Kaur, M., Fagerli, T., Temple-Bird, C., Lenel, A., & Kawohl, W. (2005a). How to Procure and Commission your Healthcare Technology, In: *World Health Organization*, September 2011, Available from: <http://www.who.int/management/procure_commission_healthcare.pdf>

Kaur, M., Fagerli, T., Temple-Bird, C., Lenel, A., & Kawohl, W. (2005b). How to Organize a System of Healthcare Technology Management, In: *World Health Organization*, September 2011, Available from: <http://www.who.int/management/organize_system_%20healthcare.pdf >

Krauss-Silva, L. (2004). Avaliação Tecnológica em Saúde: Questões Metodológicas e Operacionais. *Cadernos de Saúde Pública*, Vol. 2, No. 2, (January 2004), pp. (199-207)

Klein, G. (1997). An Overview of Naturalistic Decision Making Applications, In: *Naturalistic Decision Making*. Gary Klein & Caroline Zsambock, pp. (49-57), Lawrence Erlbaum Associates, ISBN 0-8058-1873-1, Mahway, New Jersey

Kittel, A., Marco, A. D., & Steward, H. (2002). Factors Influencing the Decision to Abandon Manual Wheelchairs for Three Individuals with Spinal Cord Injury. *Disability and Rehabilitation*, Vol. 24, No. 3, (February 15), pp. (106-114)

Joseph, A. (2007). Decision Management, *Proceedings of IEEE Military Communications Conference*, ISBN 978-1-4244-1513-7, Orlando, FL, USA, October 2007

Jørgensen, H. (2007) . Assessment of literature, In: *Health Technology Assessment,* Finn Børlum Kristensen and Helga Sigmund, pp. (57-67), National Board of Health, Retrieved from<http://www.sst.dk/publ/Publ2008/MTV/Metode/HTA_Handbook_net_final.pdf>

Mcgregor, M., & Brophy, J. M. (2005) End-user Involvement in Health Technology Assessment (HTA) Development: A way to Increase Impact. *International Journal of Technology Assessment in Health Care*, Vol. 21, No. 2 , (April 2005), pp. (263-267)

Moraes, L. (2007). Metodologia para Auxiliar na Definição de Indicadores de Desempenho para a Gestão da Tecnologia Médico-Hospitalar, In: *Domínio Público*, 21 July 2011, Available from: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=199521>

NHS - National Health Research. Policy for Training in the Safe Use of Medical Devices (2005), In: *National Institute for Health Research (NHS),* July 2011, Available from: <http://www.eastcheshire.nhs.uk/About-The-Trust/policies/T/Training-on-medical-devices.pdf>

Raymond, P. Z. (2004). Clinical Engineering, In: *Clinical Engineering Handbook,* Joseph Dyro, pp. (133-134), Academic Press, Retrieved from http://www.elsevier.com/wps/find/bookdescription.cws_home/702695/description#description

Robert, G., Greenhalgh, T. MacFarlane, & F., Peacock, R. Organisantional Factors Influencing Technology Adoption and Assimilation in the NHS: A Systematic Literature Review, In: *National Institute for Health Research (NHS),* July 2011, Available from: < http://www.sdo.nihr.ac.uk/files/project/223-final-report.pdf >

Santos, F. A. & Garcia, R. (2010). Decision Process Model to the Health Technology Incorporation, *Proceedings of 32nd Annual International Conference of the IEEE EMBS*, Buenos Aires, Argentina, August 31 - September 4, 2010

Santos, R. Souza, R. E. H. & Garcia, R. Health Care Technology Management Applied to Public Hospitals in Santa Catarina – Brazil, *Proceedings of First WHO Global Forum on Medical Devices*, Bangkok, Thailand, September 9-11, 2010

Sarwar, S. G., & Robinson, I. (2007). Benefits of and Barriers to Involving Users in Medical Device Technology Development and Evaluation. *International Journal of Technology Assessment in Health Care*, Vol. 23, No 1 , (January 2007), pp. (131-137)

Signori, M. R. & Garcia, R. Clinical Engineering and Risk Management in Healthcare Technological Process Using Architecture Framework. *Proceedings of 32nd Annual International Conference of the IEEE EMBS*. Buenos Aires, Argentina, August 31 - September 4, 2010

Simpson S., Hiller J., Gutierrez-Ibarluzea I., Kearney B., Norderhaug I., Fay AF., Packer C., Asua J., Benguria G., Blanchard S., Blozik E., Bonnevie BM., Clifford T., Eckerlund I., Galnares L., Groeneveld K., Hae Lee Robin S., Hakak N., Husereau D., Ibargoyen N., Kaila M., Künzli C., Llanos A., Luengo S., Morrison A., Mundy L., Tal O., Wallgren L., & Wallin J. (June 2009). A Toolkit for the Identification and Assessment of New and Emerging Health Technologies, In *EuroScan International Network*, July 2009, Retrieved from <http://www.euroscan.org.uk/methods>

Singer, P. A., Martin, D. K., Giacomini, M., & Pardy, L. (2000) Priority Setting for New Technologies in Medicine: Qualitative Case Study. *British Medical Journal*, Vol. 321, No. 7272, (November 2000), pp. (1316-1318)

Souza, A. F., Heringer, C., H. T., Junior, J. S., & Moll, J. R. (2010). *Gestão de Manutenção em Serviços de Saúde* (First edition). Ed. Blucher, ISBN 9788521205630, São Paulo

Sonego, F. (2007). Estudo de Métodos de Avaliação de Tecnologias em Saúde Aplicada a Equipamentos Eletromédicos, In: *Universidade Federal de Santa Catarina*, October 2011, Available from < http://www.tede.ufsc.br/teses/PEEL1174-D.pdf >

Sprague, G. R. (1988). Managing Technology Assessment and Acquisition. Healthcare *Executive*, Vol. 3, No. 6, (Nov./Dec. 1988), p.p (26-29)

Stenbæk, D. E., & Jensen, M. F. (2007). Literature searches, In: *Health Technology Assessment,* Finn Børlum Kristensen and Helga Sigmund, pp. (47-56), National Board of Health Retrieved from <http://www.sst.dk/publ/Publ2008/MTV/Metode/HTA_Handbook_net_final.pdf >

United Nations. (1948). Universal Declaration of Human Rights. Geneva. In: *United Nations*, 08 August 2011, Available from: <http://www.un.org/en/documents/udhr/index.shtml>

Velasco-Garrido M. & Busse R. (2005). Health Technology Assessment: An Introduction to Objectives, Role of Evidence, and Structure in Europe. Copenhagen, In: *World Health Organization Regional Office for Europe*, 10 October 2011, Available from: <http://www.euro.who.int/__data/assets/pdf_file/0018/90432/E87866.pdf>

Wang, B. (2009). Strategic Health Technology Incorporation, In: Synthesis Lectures on Biomedical Engineering, John D. Enderle, pp. (5-61), Morgan & Claypool Publishers Series, Retrieved from <http://www.morganclaypool.com/doi/abs/10.2200/S00216ED1V01Y200908BME032?journalCode=bme >

WHA - World Health Assembly. (2007). Health Technologies, Sixtieth World Health Assembly Agenda item 12.19. WHA60.29. In: *World Health Organization*, 15 Sebtember 2011, Available from: <http://apps.who.int/gb/ebwha/pdf_files/WHASSA_WHA60-Rec1/E/cover-intro-60-en.pdf>

WHO - World Health Organization. (2007). Everybody's Business: Strengthening Health Systems to Improve Health Outcomes: WHO's Framework for Action, In: *World Health Organization*, 12 July 2011, Available from: <http://www.who.int/healthsystems/strategy/everybodys_business.pdf>

WHO - World Health Organization. (2008). World Health Report 2008: Primary Health Care–Now More than Ever. In: *World Health Organization*, 10 July 2011, Available from: <http://www.who.int/whr/2008/whr08_en.pdf>

WHO - World Health Organization. (2009). Systems Thinking for Health Systems Strengthening, In: *World Health Organization*, 14 July 2011, Available from: <http://whqlibdoc.who.int/publications/2009/9789241563895_eng.pdf>

WHO - World Health Organization. (2011a). Medical Device Technical Series. Procurement Process Resource Guide, In: *World Health Organization*, 16 July 2011, Available from: <http://whqlibdoc.who.int/publications/2011/9789241501378_eng.pdf>

WHO - World Health Organization. (2011b). Medical Device Technical Series. Development of Medical Device Policies, In: *World Health Organization*, 16 July 2011, Available from: < http://whqlibdoc.who.int/publications/2011/9789241501637_eng.pdf>

WHO - World Health Organization. (2011c). Medical Device Technical Series. Medical Equipment Maintenance Programme Overview, In: *World Health Organization*, 17 July 2011, Available from: <http://whqlibdoc.who.int/publications/2011/9789241501538_eng.pdf >

WHO - World Health Organization. (2011d). Medical Device Technical Series. Needs Assessment for Medical Devices, In: *World Health Organization*, 23 July 2011, Available from: <http://whqlibdoc.who.int/publications/2011/9789241501385_eng.pdf >

Woodside, A. G., Breaux, R., & Briguglio, E. (1998). Testing Care-Giver Acceptance of New Syringe Technologies. *International Journal of Technology Management*, Vol. 15, No. 3/4/5, pp (446-457)

Zsambock, C. (1997). Naturalistic Decision Making Where Are We Now?, *In: Naturalistic Decision Making*, Gary Klein & Caroline Zsambock, pp (3-28), Lawrence Erlbaum Associates, ISBN 0-8058-1874-X, Mahway, New Jersey

14

Policy and Management of Medical Devices for the Public Health Care Sector in Benin

P. Th. Houngbo[1,3], G. J. v. d. Wilt[3], D. Medenou[2],
L. Y. Dakpanon[1,2], J. Bunders[3] and J. Ruitenberg[3]
[1]*Ministry of Health,*
[2]*Polytechnic School, University of Abomey-Calavi,*
[3]*Athena Institute, Vrije Universiteit, Amsterdam,*
[1,2]*Republic of Benin*
[3]*The Netherlands*

1. Introduction

Health technology, according to WHO is the application of organized knowledge and skills in the form of devices, medicine, vaccines, procedures and systems development to solve a health problem and improve quality of lives[4]. When used in this paper, the term healthcare technology means the different types of devices or equipment used in health facilities. Its encompasses: medical equipment for clinical use; hospital furniture; vehicles; service Supplies; plant; communication equipment; fire fighting equipment; fixtures built into the building; office equipment; office furniture; training equipment, walking aids and workshop equipment.

Healthcare technologies offer many benefits and have greatly enhanced the ability of health professionals to prevent, diagnose and treat diseases[11]. They are one of the essential elements for the delivery of health services. The use of technology in health care systems in developing and transition countries faces a great number of difficulties. Since about 95% of the healthcare technology used in these countries is imported[30]; mismatches occur because the technology development process has not usually considered the needs and realities of the target environments. These mismatches in the technology transfer process to countries with financial and technical constraints are often of great significance. Thus, in Benin, medical devices and equipment represent a significant proportion of national health care expenditure. Each year, more than 10,600,000 US$, (about 20%)[20] of the national health budget, are spent on procurement of medical devices and equipment for healthcare facilities. Despite this great amount of money spent each year on an ever-increasing array of medical devices and equipment, not enough attention is paid to the equipment use and maintenance. Management of medical devices is not yet recognised as an integral part of public health policy. Planning, follow up and maintenance of the equipment are inefficient and ineffective [12, 13, 14, 15, 16, 17, 18, 19, 20, and 21].

This study, supported by the *Netherlands Organisation for International Cooperation in Higher Education* (NUFFIC) from 2007 was conducted in Benin Ministry of Health (MoH) and at the University of Abomey-Calavi in collaboration with the Athena Institute, Vrije Universiteit Amsterdam from 2006-2008 aimed to identify factors appearing between 1998 and 2008 that adversely affected the healthcare technology management cycle i.e., planning, budgeting, selection, procurement, distribution, installation, training, operation, maintenance and disposal of medical devices. The results will allow to identify the key factors of mismanagement and critical maintenance system of medical devices in Benin and to formulate recommendations to improve the system. The first part of this paper gives background information on the country, its health system and an overview of its healthcare technology management state. The second part describes the methods and materials used and the third part presents the results, followed by discussion, comments and recommendations in the final section.

2. Background information

2.1 Benin: The country

Located on the West coast of Africa, the Republic of Benin is small (114,763 square kilometers), with a coastline on the Gulf of Guinea nestled between Nigeria, Niger, Burkina Faso, and Togo (Figure 1). The population, estimated at 7,839,914 in 2006, includes a multitude of ethnic and linguistic groups. Benin remains one of the world's least developed countries and has been ranked 163 of 177 on the United Nations Human Development Index (2005). Demographic and health indicators are given below (Table 1).

Fig. 1. Map of Benin (*Source:* USAID, 2006)

Indicators	
Population in 2006	7,839,914
Human Development Index	0.437
Country rank	163/177
GPD per capita (Purchasing Power Parity US$)	1,141
Life expectancy at birth (years)	55.4
Public expenditure on health (% of GPD) in 2004	4.5
Health expenditure per capita (PPP US$) in 2004	40
Infant mortality rate per 1,000 live births	67
Maternal mortality ratio per 100,000 live births	474
HIV/AIDS prevalence (%)	2.0
Adult literacy rate (% ages 15 and older)	34.7

Sources: 1. Human Development Reports: 2007/2008;
2. Benin Demographic and Health Survey 2006;
3. Benin Health Statistics Directory 2006.

Table 1. Selected demographic and health indicators of Benin

2.2 The health system

The public healthcare system of the country has been reorganized according to the decentralization policy and consists of three levels: **central** with the National Referral Hospital (> 600 beds), **intermediate** with five Province or Departmental Hospitals (>100 beds) and **peripheral** with thirty four Health Zones, twenty seven fairly functional Zone Hospitals (> 46 beds), seventy seven Communal Health Centers, four hundred eighty seven Arrondissement Health Centers and many Village Health Units and other private health facilities. Apart from that, the health system also has the following public hospitals: the Mother and Child Hospital, many detection and treatment centers for tuberculosis and leprosy, the National Hospital for Psychiatry, the National Hospital for Gerontology, two Buruli Ulceration Treatment Centers[12] etc…

2.3 Healthcare technology management and maintenance

Healthcare technology management and maintenance remains one of the main challenges of the developing countries healthcare systems in general and, of Benin particularly. Thus, although many financial resources are used for procurement of devices, not enough attention is paid to their future. While some of the equipment were donated, a significant portion was purchased with loans provided by bilateral and multilateral agencies and will have to be paid back with great sacrifice[26]. One of the root causes of the equipment idleness is the lack of effective management. It is important to point out that despite the several initiatives undertaken by the ministry of health to improve the *healthcare technology management cycle* no significant changes have been noticed[13, 14, 15, 16 and 17].

Many facilities, especially Zone Hospitals, continue to lack the basic technologies they need to provide quality care to the patients, because equipment is unavailable, inoperative, misused or inappropriate. The situation is most severe in the Communal and Arrondissement health facilities far from the first referral hospitals. This has far-reaching implications for the prevention and treatment of disease and disability and often leads to a waste of scarce resources.

3. Materials and methods

The study was carried out in the MoH, 321 healthcare facilities of the southern part of the country, the Ministry of Economy and Finance, some representatives of external support agencies in Benin and ten accredited suppliers of medical device companies. It consisted of surveys undertaken in 2006 and 2007 and of desk research (content analysis) based on 1998 to 2008 procurement collected data. It aimed to determine the factors that adversely affect the healthcare technology management cycle (planning, budgeting, selection, procurement, distribution, installation, training, operation, maintenance and disposal of medical devices) in Benin.

3.1 Desk research and short survey

This study focused on the procurement management of medical devices in the Republic of Benin and aimed to identify the main weak points in the procurement management system of medical devices from 1998 to 2008. It was based on data collected from documents (such as national procurement magazines and health equipment public procurement and bidding contracts from the Ministries of Health and Economy and Finances), and on interviews and informal discussions with ten local accredited suppliers of medical devices in Benin.

A comparative study was done concerning the selling prices of ten medical devices procured by Benin MoH further to international tenders. The steps were i) Ten medical devices were selected from the available essential medical device list. ii) Their mean reference selling prices (based on their specifications) were determined from 10 local medical device accredited suppliers based on the prices the devices were sold to the private health facilities. iii) The mean prices at which the same devices were sold to the Ministry of Health following open tenders public procurement were identified, in three periods: 1998 to 1999; 2001 to 2004 and 2005 to 2008 when the procurement evaluation process has been changed and improved. iv) The mean prices at which they were sold to the MoH were compared to the ad hoc mean reference selling prices provided by the private healthcare facilities and/or from the local suppliers' price list for private facilities.

3.2 Surveys

Two surveys were carried out in 321 healthcare facilities of the six southern departments (provinces). The first, entitled "management and maintenance of healthcare technology", was conducted in 2006 in 11 health centers and hospitals. It aimed to identify the weaknesses in the healthcare technology management and maintenance system in order to make recommendations for its improvement. Data were collected through observational visits, interviews and questionnaires. The second, entitled "healthcare technology assessment in the southern Benin public healthcare facilities" was carried out in 310 health centers and hospitals in 2006 and 2007. The first objective was to determine the extent of disparity between what medical devices/equipment were planned and what was actually available in each selected health facility to facilitate procurement for the poorly equipped health facilities of the essential medical devices. The second objective

was to identify weaknesses in the whole Benin healthcare technology management cycle. Data were collected through observational visits and reading reports, interviews, and questionnaires (inventory sheets). The steps were i) Equipment inventory was done at all the public healthcare facilities in southern Benin; ii) Healthcare equipment in these facilities were compared to the MoH available Essential Medical Device List of each health facility level iii) The needs assessment of each healthcare facility was done using a pilot asset assessment software. Finally, interviews were held with a range of stakeholders including policy makers of the MoH, healthcare facility managers, equipment users (physicians, nurses, midwives, lab technicians, X-ray machine technician ….) and, maintenance technicians.

4. Results

The results of the study are summarised in tables 2 to 6 and figure2. Tables 2, 3 and 4 show the mean ad hoc reference selling prices of selected medical devices in comparison with the prices the same devices were sold to the Ministry of Health from 1998 to 1999, 2001 to 2004 and 2005 to 2008. Table 5 and figure 2 show the trends of [MoH device acquisition prices/Ad hoc device reference selling prices] ratio during the three periods of years. The ten equipment studied were: 1) blood pressure device 2) spectrophotometer 3) electric suction unit 4-) Electrocardio-graph 5) X-ray apparatus 6) hot air sterilizer 7) autoclave 8) ventilator 9) anaesthesia system and 10) blood bank refrigerator.

The letter X that may be a, b, c, d, e, f, g, h, i or j represents respectively the "ad hoc reference prices" (the private healthcare facilities device acquisition prices) of each device in local currency. The letter Y that may be A, B, C, D, E, F, G, H, I or J are respectively the MoH same device acquisition prices through public procurement. Table 6 presents the findings of the two surveys and shows the factors affecting the healthcare technology management cycle in 321 health centers and hospitals in southern Benin. The factors were grouped (but not ranked) in sixth categories which were respectively maintenance and repair; distribution; use; technology assessment; policy, planning and budgeting; and procurement.

The key factors that have been identified so far include the high acquisition costs; the lack of insight of the government on medical device market prices, the lack of capacity to monitor reasonable prices from suppliers, the lack of insight into the cost/performance ratio of various brands of medical devices, an unequal distribution of devices among health care facilities, an unbalanced allocation of resources to acquisition of devices compared to infrastructure, and maintenance. Other key factors identified included the insufficiency of human resources with appropriate capacity to manage equipment, the unavailability of spare parts, and the lack of an annual maintenance budget. In a nutshell, the lack of policy and management tools like "the up to date essential medical devices list and "the reference prices list for essential medical devices" to support the implementation of the existing policy. The latter allows health sector authorities to monitor financial diversions occurring in public procurement contract awards, while the former serves as a reference tool to assess availability of fully operational devices at different hierarchical levels of healthcare facilities.

4.1 Findings of the desk research and short survey

Devices	Ad hoc reference prices (Private healthcare facilities same device acquisition prices) X	MoH device acquisition prices through public procurement Y	Y/X Ratio
1. Blood pressure device	a	A= 3.13a	3.13
2. Spectrophotometer	b	B= 4.00b	4.00
3. Electric suction unit	c	C= 2.85c	2.85
4. Electrocardio-graph	d	D= 2.38d	2.38
5. X-ray apparatus	e	E= 2.12e	2.12
6. Hot air sterilizer	f	F= 2.38f	2.38
7. Autoclave	g	G= 2.22g	2.22
8. Ventilator	h	H= 2.85h	2.85
9. Anaesthesia system	i	I= 3.33i	3.33
10.Blood bank refrigerator	j	J= 2.32j	2.32
Arithmetic mean of Y/X ratio =			**2.75**

Table 2. Comparison of the mean ad hoc reference prices of medical devices to the Ministry of Health same device acquisition prices, 1998 to 1999.

Devices	Ad hoc reference prices (Private healthcare facilities same device acquisition prices) X	MoH device acquisition prices through public procurement Y	Y/X Ratio
1. Blood pressure device	a	A= 6.66a	6.66
2. Spectrophotometer	b	B= 5.55b	5.55
3. Electric suction unit	c	C= 6.66c	6.66
4. Electrocardio-graph	d	D= 4.00d	4.00
5. X-ray apparatus	e	E= 3.13e	3.13
6. Hot air sterilizer	f	F= 2.63f	2.63
7. Autoclave	g	G= 5.00g	5.00
8. Ventilator	h	H= 4.00h	4.00
9. Anaesthesia system	i	I = 4.34i	4.34
10.Blood bank refrigerator	j	J = 3.03j	3.03
Arithmetic mean of Y/X ratio =			**4.50**

Table 3. Comparison of the mean ad hoc reference prices of medical devices to the Ministry of Health same device acquisition prices, 2001 to 2004.

Devices	Ad hoc reference prices (Private healthcare facilities same device acquisition prices) X	MoH device acquisition prices through public procurement Y	Y/X Ratio
1. Blood pressure device	a	A= 2.38a	2.38
2. Spectrophotometer	b	B= 3.33b	3.33
3. Electric suction unit	c	C= 2.38c	2.38
4. Electrocardio-graph	d	D= 2.25d	2.25
5. X-ray apparatus	e	E= 2.22e	2.22
6. Hot air sterilizer	f	F= 2.38f	2.38
7. Autoclave	g	G= 2.38g	2.38
8. Ventilator	h	H= 2.63h	2.63
9. Anaesthesia system	i	I= 2.85i	2.85
10.Blood bank refrigerator	j	J= 2.08j	2.08
Arithmetic mean of Y/X ratio =			**2.48**

Table 4. Comparison of the mean ad hoc reference prices of medical devices to the Ministry of Health same device acquisition prices, 2005 to 2008.

Devices	Y/X ratio (1998-1999)	Y/X ratio (2001-2004)	Y/X ratio (2005-2008)
1. Blood pressure device	3.13	6.66	2.38
2. Spectrophotometer	4.00	5.55	3.33
3. Electric suction unit	2.85	6.66	2.38
4. Electrocardio-graph	2.38	4.00	2.25
5. X-ray apparatus	2.12	3.13	2.22
6. Hot air sterilizer	2.38	2.63	2.38
7. Autoclave	2.22	5.00	2.38
8. Ventilator	2.85	4.00	2.63
9. Anaesthesia system	3.33	4.34	2.85
10.Blood bank refrigerator	2.32	3.03	2.08
Arithmetic mean of the three ratios=	**2.75**	**4.50**	**2.48**

Table 5. Trend of the [MoH device acquisition price/Ad hoc device reference prices] ratio during the three periods of years: 1998-1999; 2000-2004 and 2005-2008.

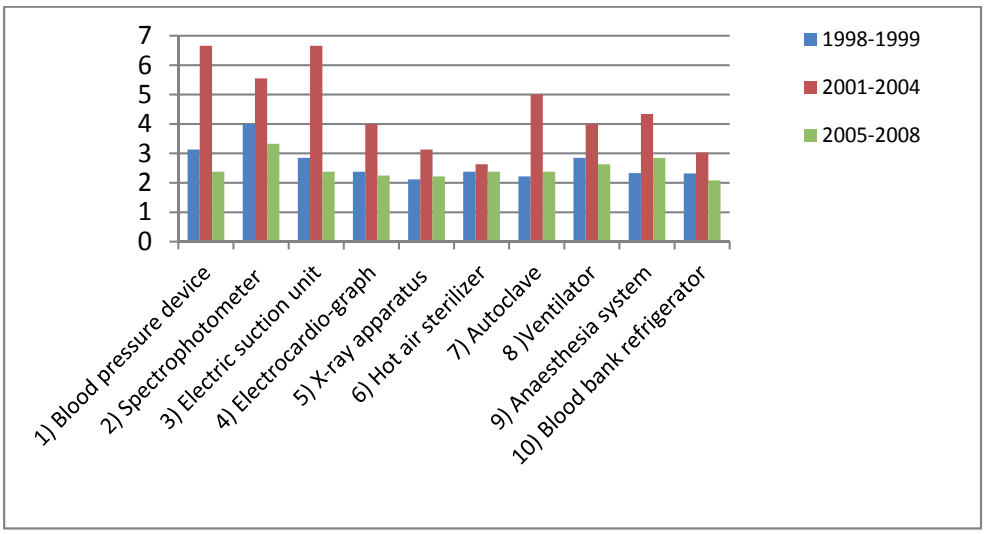


Fig. 2. Comparative graphs of the MoH selected medical device acquisition prices during the three periods of years:1998-1999; 2001-2004 and 2005-2008.

4.2 Finding of the surveys 1 and 2

1. Management and maintenance of healthcare technology
2. Healthcare technology assessment in the southern Benin public health facilities.

Categories	Factors
I	**Maintenance and repair** ■ Unavailability of many equipment. ■ Lack of maintenance after-sales service. ■ Lack of availability of equipment spare parts ■ Lack of equipment service manuals. ■ Lack of preventive maintenance of the equipment. ■ Lack of resources (financial, material and Human resources) for implementation of maintenance activities.
II	**Distribution** ■ Unequal and inappropriate distribution of devices within the healthcare facilities. ■ Inappropriate technology for the site, size, capacity of healthcare facilities.
III	**Use** ■ Lack of continued training and misuse of equipment ■ Lack of equipment user manuals
IV	**Technology assessment** ■ Lack of equipment assessment ■ Lack of implementation of asset management tools (software)
V	**Policy, planning and budgeting** ■ Lack of awareness of healthcare technology management issues. ■ Lack of implementation of the existing healthcare technology policy manual. ■ Lack of effective and efficient maintenance planning. ■ Lack of sufficient annual maintenance budget allocation.
VI	**Procurement** ■ High medical devices acquisition costs through public procurement. ■ Insufficient involvement of the equipment end users into the device acquisition process. ■ Lack of information on the device acquisition costs in the healthcare facilities for assess replacement planning. ■ Ineffective planning and inappropriate devices procurement (Acquisition of heavy and specific equipment without effective planning of the architectural and technical installation requirements. ■ Multiplicity of makes due to the multiple acquisition sources.

Table 6. Factors affecting the healthcare technology management cycle in 321 health centers and hospitals in southern Benin.

5. Discussion and recommendations

Goods acquisition, especially healthcare technology, represents an important part of any health budget and need to be looked with close attention. Through the results shown in Tables 2, 3, 4, 5 and, figure 2, it is clearly seen that, independently of the procurement years, the device acquisition prices by the MoH remain higher than the private healthcare facilities same device acquisition prices. Although the Benin first Goods and Services Procurement Code was implemented during the years 2001 to 2004 and has also been amended in 2004 and be implemented from 2005 to 2008, no significant improvements were found regarding the higher prices of medical equipment paid by the MoH. One can notice that the MoH pays too much for medical devices acquisition through public procurement and this was at its worst in 2001-2004. When analysing year by year available data of this period it was found that the highest acquisition prices were critical in 2003 and 2004. It is important to deeply understand the real reasons that underlie this phenomenon. Many hypotheses could be drawn to explain this fact but, it will be more interesting to increase the sample size (>10 medical devices) of the study for more reliability. The internal and external validities of the findings could be improved if a quasi-experimental study was designed. Thus, widely surveys will be conducted in the next papers with more representative sample size and strong method as controlled interrupted time series based on segmented regression analysis

to infirm or to confirm the present findings and also to understand the true reasons of the ineffective management of healthcare equipment in Benin.

The Ministry of Health still needs a national public procurement policy and management tool like a reference prices list of the most widely used devices to overcome and to master the increasing and unreasonable medical device prices. It is normal to have the device acquisition costs paid by the government a bit higher than the reference set prices because of financial and administrative fees involved when the suppliers submit tenders. It is acceptable and reasonable to have the average device selling prices comprised between **1.1 to 1.2 times** higher than the ad hoc reference prices. But, when the device selling prices offers by a supplier are more than that, they could be considered as *outbidding*. It is thus urgent for the Benin government especially the MoH to have an insight on that fact, to encourage the development of policies and laws regarding a reference price lists document of medical devices. The availability of the reference prices of the essential medical devices will allow the health sector authorities to monitor the usual financial diversion occurring during the procurement management activities. It is expected that once this document becomes available, the MoH could buy value-based pricing equipment each year and save a lot of money that can be used to improve the health of Benin population through other investments.

The results of the two surveys: i) "management and maintenance of healthcare technology" and ii) "healthcare technology assessment in the southern Benin public healthcare facilities" have revealed many weaknesses in the Benin health system through its healthcare technology management cycle. The results show failures in each link of the cycle (planning, budgeting, selection, procurement, distribution, installation, training, operation, maintenance and disposal of medical devices) resulting in low overall community health effectiveness. It is necessary to point out that the findings of the two surveys, i.e. the factors affecting healthcare technology management were only grouped (but not ranked) in sixth categories. The ranking of the factor categories (I, II, III, IV, V and VI) in order to set up priority actions will be discussed in the next paper.

As recommendations, twenty actions need to be taken by the government to overcome this situation in order to achieve its goal to improve the quality of/and access to health services that taking into account the poor and indigent. It is thus urgent to develop and implement a good medical device national policy which can include the following: i)An improved national list of essential medical devices and equipment based on evidence from the studies; ii) A national policy and plan for medical devices; iii) A national functional regulation authority in medical device empowered with legislation; iv) A document on assessment of medical device needs; v) National regulations based on ISO standards or WHO specifications; vi) National procurement procedure; vii) National policy for acceptance of donations; viii) Negotiated pricing list of each item of equipment; ix) National guide for management and use of medical devices; x) An inventory of suppliers and medical in use; xi) The cost of all the equipment of each level of Benin health facility related to the cost of infrastructure; xii) The service life span of each medical device or equipment in use in Benin health care facility or hospital in order to plan the replacement at a systematic time; xiii) The list of medical devices which have the highest risk; xiv) The spare parts which have the highest failure rate in order to plan their procurement; xv) The list of critical equipment and instrument affected by the electrical power outages and power anomalies in Benin hospitals; xvi) Good software based planning and management tools for management and

maintenance of medical devices; xvii) A post-market surveillance/vigilance system for alerts, notifications and recalls; xviii) A national budget for devices, using costing, budgeting and financing; xix) Standard operating procedures and best practices that cover every stage in the life span of medical devices; xx) Creation of an independent Direction of Healthcare Technology Management and maintenance within the Ministry of Health.

The following Healthcare Technology Management Cycle (Figure 3) could be used as a framework for health equipment management in developing country, providing a guideline for the necessary regulations and systems.

The Healthcare Technology Management Cycle[11]**:**An example of a framework for health equipment management in developing country.

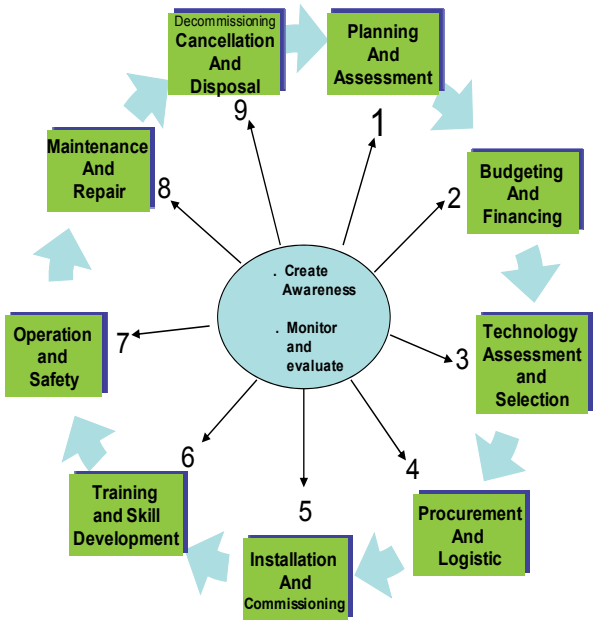


Fig. 3. The Healthcare Technology Management Cycle

6. Conclusion

Management and maintenance of healthcare technology in developing countries especially in the poor sub-Saharan Africa countries, remain a challenge. From the planning to the disposal of the devices many actions need to be undertaken to improve the Healthcare Technology Management Cycle. The achievements in the public healthcare sector depend on the full involvement of each stakeholder, but the main responsibility is still that of the governments. They need the political willingness and commitment to recognize management and maintenance of devices as an integral part of public health policy in order improve the quality and access to healthcare in each country.

7. Acknowledgement

We would like to thank very much The Netherlands Organization for International Cooperation in Higher Education (NUFFIC) that grants fellowship for the main investigator to undertake PhD research. Thanks are also due to the technical officers of the Ministry of Health for their collaboration. The authors are grateful to Prof E. P. Wright for her fruitful comments.

8. References

[1] Bloom, G., Temple-Bird, C.: *Medical Equipment in Sub-Saharan Africa: A Framework for Policy-Formulation*. IDS research Report No.19, WHO/SHS/NHP/90.6, WHO, Geneva, 1990.

[2] Benin Tourisme: *Benin; histoire [online]. 2007 Aug 25* Available from: URL: www.benintourisme. com.

[3] Department of Health, Republic of South Africa: *A Framework for Health Technology Policies*.

[4] Fahlgren B.: *Access to effective medical technology in Developing Countries-what role for WHO?* WHO Geneva 2004.

[5] Goodman, C.S. and Ahn, R.: *Methodological Approaches Used in Health care Technology Assessment*. NICHSR, USA 2004.

[6] Gouvernement du Benin. *Développement économique* Available from: URL: www.gouv.bj.

[7] Guinand C.: *Maintenance biomédicale. Zones sanitaires appuyées par le PBA-SSP. Evaluation et suivi des activités des techniciens.* Décembre 2000. Cotonou, 2000.

[8] Heimann, P., Poluta, M.A.: *Health Technology Management in Sub-Saharan Region as a Pre-requisite for Optimising the Donor Aid Intervention Process.* (In press) WHO, ARA, Geneva, 1997.

[9] Institut National pour les Statistique et L'analyse Economique : *Enquête Démographique et de Santé,* Cotonou 2006.

[10] Issakovov, A.: *Service and Maintenance in Developing Countries, pp. 21-28 in:* Medical Devices: International Perspectives on Health and Safety. Ed. Van Gruting C.W.D. Elsevier, Amsterdam, 1994.

[11] Keller J.P.JR., and Walker S.: *Best Practices for Medical Technology Management: A U.S. Air Force-ECRI Collaboration*: Advances in Patient Safety: Vol. 4. pp 45-55, USA, 2004.

[12] Ministère de la Santé: *Annuaire des statistiques sanitaires de la République du Bénin. Edition 2006.*

[13] Ministère de la Santé Publique: *Atelier National d'Orientation des Politiques et Stratégies Nationales de Maintenance Hospitalière en République du Bénin à Possotomè du 21 au 23 février 2000 . Possotomè* 2000.

[14] Ministère de la Santé Publique de la République du Bénin : *Avant-projet de politique et stratégies de maintenance des infrastructures et équipements médicaux au Bénin, Février 2000 . Possotomè* 2000.

[15] Ministère de la Santé de la République du Bénin: *Etude d'évaluation de la situation actualisée des plateaux techniques des formations sanitaires publiques par niveau de soin et vérification de leur conformité aux normes dans les six (06) départements du sud Bénin,* Bénin 2006.

[16] Ministère de la Santé Publique de la République du Bénin : *Etude sur l'élaboration d'un système décentralisé de maintenance hospitalière (30 mars au 15 avril 1995).* Cotonou 1995

[17] Ministère de la Santé de la République du Bénin : *Politique de maintenance des infrastructures, des équipements médico-techniques et du parc automobile en République du Bénin.* Cotonou 2002

[18] Ministère de la Santé Publique de la République du Bénin : *Politiques et stratégies nationales de développement du secteur santé (1997-2001)* . Cotonou 1998.

[19] Ministère de la Santé Publique de la République du Bénin : *Politiques et Stratégies Nationales de Maintenance Hospitalière en République du Bénin (2001-2005).* Cotonou 2000.

[20] Ministère de la Santé de la République du Bénin : *Rapport de la mission d'expertise thématique en gestion et maintenance des équipements et infrastructures de la santé*, Bénin 2005

[21] Ministère de la Santé de la République du Bénin: *Recueil d'informations de la Cellule de Passa-tion des Marchés*, Cotonou, 2006.

[22] Projet Bénino-Allemand des Soins de Santé Primaire (PBA-SSP): *Guide d'entretien du Matériel des CSSP et CCS du Projet Bénino-Allemand des Soins de Santé Primaires* . Cotonou 1997

[23] South African Medical Research Council: *Executive Report of the Regional Workshop on Health care Technology in Sub-Saharan Region*, Somerset West, South Africa. April 1994

[24] USAID : Rapid assessment of the health system in Benin. Benin 2002

[25] USAID: *Country background*; Benin. Available from: URL: www.usaid.gov.

[26] Wang, B.: *A framework for Health Equipment Management in Developing Countries*. Hospital Engineering and Facilities Management 2003

[27] World Health Organization: *Medical Device Regulations: Global Overview and Regulating Principle.* WHO, ISBN 92 4 154618.

[28] World Health Organization: *Regulation Challenges (Medical device regulation: a framework).* WHO Drug Information Vol 17, No. 4, 2003.

[29] World Health Organization: *Experts on Healthcare Technology in the Developing World Meet at Savoy Place (Managing Health care Technology)* WHO. 2, 2002.

[30] World Health Organization: *Essential Health Technologies: Strategy 2004- 2007*. WHO. Draft for comments by member States. 11, 2003.

[31] World Health Organization.: *Global Harmonization Task Force (Working Toward Harmonization in Medical Device Regulation Full Document)*. Geneva 2004.

Section 4

Global Health

15

Non-Communicable Diseases in the Global Health Agenda

Julio Frenk[1], Octavio Gómez-Dantés[2] and Felicia M. Knaul[3]
[1]Harvard School of Public Health, Boston, MA,
[2]Center for Health Systems Research, National Institute of Public Health,
[3]Harvard Global Health Equity Initiative and Harvard Medical School, Boston, MA,
[1,3]USA
[2]Mexico

1. Introduction

For a long time non-communicable diseases (NCDs) have been a major cause of death and disability worldwide. However, the profile of this health challenge is changing: Having dominated the epidemiologic contour of high-income countries in the 20th century, it is now increasingly affecting the developing regions of our planet. Unless we start implementing measures to reduce the burden of NCDs in low- and middle-income countries, the pressure on their health systems will be unbearable and will limit the prospects for economic development.[1]

In this chapter we discuss the need to confront this emerging challenge through a change in the orientation of global health. The central message is that it is necessary to incorporate NCDs into the global agenda and deploy comprehensive strategies in developing countries to address them. Such strategies should include both prevention services and cost-effective treatments.

In the first part of the chapter we discuss the present situation of NCDs in low- and middle-income countries, with emphasis on cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes, along with a major risk related to most of them, obesity. Part two is devoted to the discussion of four myths that have hindered the incorporation of NCDs to the global health agenda and a set of proposals to strengthen the battle against them, using as an example several initiatives implemented in Mexico as part of a comprehensive health reform. The chapter concludes with a call to mobilize international collective action in the pursuit of shared goals around NCDs.

2. The global burden of NCDs

During the past half-century the world witnessed a fundamental transformation in the field of health: a shift in the dominant patterns of disease and death towards higher age groups and towards chronic conditions.

Improvements in nutrition, access to water and sanitation, and expanded coverage of public health interventions such as immunizations and oral rehydration therapy reduced the burden of disease attributed to under-nutrition, common infections and reproductive

problems, and produced major gains in child survival beyond age 5. Recently, the expansion of the global coverage of immunizations, for example, produced a 74% drop of measles deaths between 2000 and 2007.[2] The number of global deaths due to malaria declined from almost one million in 2000 to 780 thousand in 2009.[3] Annual maternal deaths also fell from more than a half a million in 1980 to less than 350 thousand in 2008.[4]

The gains made against infectious diseases and advances in child survival rendered huge improvements in life expectancy. In fact, during the 20th century the world as a whole experienced a larger gain in life expectancy than in all the previously accumulated history of humankind. Life expectancy was only 30 years in 1900. By 1985 it had more than doubled to 62 years. In 2010 the average estimate for the world reached 70 years, but with huge regional differences, ranging from 83 years in Japan to scarcely 47 years in Zimbabwe.[5]

Today growing proportions of the world population are living long enough to experience the effects of the exposure to health risks related to modern living such as lack of physical activity, consumption of unhealthy diets and products (tobacco, alcohol and illicit drugs), stress and social isolation, all of which increased the prevalence of NCDs to the point of turning them into the leading cause of death worldwide. According to a recent World Health Organization (WHO) report, two thirds (36 million) of the total annual deaths are attributed to these diseases and 80% of them occur in low- and middle-income countries.[6]

The most common NCDs are cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes. Heart diseases are the main cause of death worldwide. They produce 17 million deaths annually, 80% of which occur in low- and middle-income countries.[6] In fact, deaths due to heart diseases today are more numerous in China and India than in all the developed world.

Cancer is another major challenge. According to WHO, there are 7.6 million cancer deaths annually worldwide, which represent around 21 percent of all NCD deaths.[6] Two thirds of them occur in low- and middle income countries.

The most common cancers among women in developing nations are breast, cervical, stomach, lung, and colorectal cancer.[7] Every year more than half a million new cases of breast cancer occur in this part of the world.[8] In Latin America, Uruguay (83 per 100,000 women) and Argentina (75 per 100,000 million) have already reached breast cancer incidence rates similar to that of Canada (96 per 100,000 women), which is one of the highest in the world.[9] Cervical cancer, which has become a rare disease in rich nations, produces more than 200,000 deaths annually in developing countries.[10]

Among men, the most common neoplasms in developing nations are lung, stomach, liver, esophageal, and colorectal cancer.[7] While rich countries are witnessing a decline in new cases of lung cancer as a result of broad anti-smoking campaigns, many low- and middle-income nations are experiencing the opposite trend. Liver cancer is also increasing among men in poor countries. More than 80% of the new cases of this disease occur in developing nations, with Sub-Saharan Africa and Southeast Asia showing the highest rates worldwide. It comes as no surprise to find out that in these same regions hepatitis B virus infection is endemic.

The third major group of NCDs is formed by chronic respiratory diseases, including asthma and chronic obstructive pulmonary disease, which produce 4.2 million deaths annually.[6]

Diabetes is the fourth major non-communicable challenge. The number of adults with this disease has doubled in the past three decades, from 153 million in 1980 to 347 million in 2008.[11] This disease produces 1.3 million deaths annually, more than 80% of them in developing regions.[6] To this we should add its morbidity impacts, since diabetes is the leading cause of renal failure, limb amputation, and visual impairment and blindness. This imposes huge economic burdens on individuals, households, health care systems, and national economies. According to a report of the International Diabetes Federation, total expenditure on diabetes reached 376 billion dollars in 2010 and is projected to exceed 490 billion dollars by 2030.[12]

Obesity is closely related to the increasing prevalence of cardiovascular diseases, several forms of cancer, and diabetes. According to a paper recently published in *The Lancet*, there are 1.46 billion overweight adults globally; 495 million of them obese.[13] Among other factors, this is the result of recent changes in the global food system which is producing increasing amounts of affordable processed food.[14] Obesity levels range from 3% in Japan to around 80% in some of the islands of the South Pacific. Children are being increasingly affected. A report of the US Institute of Medicine indicates that 20% of American children between the ages of 2 years and 5 years are overweight or obese.[15] Figures of the latest National Health and Nutrition Survey in Mexico indicate that the prevalence of obesity among children 5 to 12 years old increased from 6% to 10% between 1999 and 2006.[16] In the developing world this epidemic first affected the affluent middle-aged adults in urban settings, but it is now spreading to rural areas and indigenous populations, affecting younger age groups, and rapidly turning obesity into a disease of the poor.

If the present trends continue, by 2050 more than 50% of the world population could be clinically obese and national health systems would be overburdened by the demands associated to this health risk.[17] Withrow and colleagues estimated that obese individuals have medical costs 30% higher than those with normal weight.[13]

The shift of the burden of disease in developing countries towards chronic conditions is demanding the design and implementation of new local health strategies, but it is also calling for changes in the contents and orientation of the global health agenda. In the following section we will discuss four myths that have delayed the incorporation of NCDs to the global agenda and a set of proposals to successfully address these emerging challenges.

3. Overcoming the barriers to incorporate NCDs to the global agenda

During the 20th century international health was mostly involved in the control of communicable diseases, which were supposed to be characteristic of developing countries and mostly controlled in the developed world. NCDs, in contrast, had a low profile in the global health agenda, under the belief that they would be limited for quite a long time to high-income countries. In those days there was also a general consensus around the idea that infections and NCDs were biologically un-related.

Reality proved to be more complex. Infections never disappeared from the developed world. AIDS and antibiotic resistance have been strong remainders of the danger of lowering the guard against communicable diseases. As shown in the previous section, NCDs are increasingly dominating the health profile of the developing world. Finally, many ailments originally classified as non-communicable have now been found to have an

infectious cause. According to WHO, one fifth of all cancers worldwide are caused by chronic infections produced by agents such as HIV, HPV, and hepatitis B virus.[18] Bacterial, viral, and parasitic infections also underlie other NCDs, such as rheumatic heart disease, Chagas cardiomyopathy, and peptic ulcer.

To make matters more complex, many NCDs can literally be transmitted through genetic, epigenetic, and social networking mechanisms. The former Director General of WHO, Gro Harlem Brundtland, used to talk about "communicated diseases," which may be non-communicable in the epidemiologic sense of the word, but are transmitted through advertising and sponsorship of unhealthy products such as tobacco, junk food, and soft drinks.[19]

If we are to successfully meet the NCD challenge, we must overcome the four following myths, which have been identified in the work of the Global Task Force on Expanding Access to Cancer Care.

Myth #1: NCDs are not a major problem in developing countries. As shown above, a solid body of evidence has documented the rising importance of NCDs. According to the WHO *Global Status Report on NCDs*, nearly 80% of NCD deaths occur in low- and middle-income countries, and even in African countries they will exceed communicable, maternal, perinatal, and nutritional diseases as the most common causes of death by 2030.[6]

Myth #2: *Even if the NCDs are important, there is very little that developing nations can do to address them.* Actually, we have at our disposal cost-effective interventions for the majority of NCDs common in developing regions, and we should deploy them alongside preventive strategies in what has been called the full cycle of care.[20]

Myth #3: Even if there are effective interventions, developing countries cannot afford them. Several experiences show that it is feasible to mobilize both global and national resources in a fiscally responsible way to greatly expand access to comprehensive services for NCDs.

Myth #4: Responding to the challenge of NCDs would distract attention from other more urgent priorities, mainly the health-related Millennium Development Goals (MDGs). This myth is especially pernicious because it tends to polarize the global health community in a zero-sum, competitive mentality. Instead, we should look for synergies among disease-specific programs and strengthen health systems so that they can address the multiple, diverse, and complex needs of real people. A solid health system will be able to meet the needs related both to the unfinished agenda of common infections and to the emerging burden of NCDs.

These four myths sound very familiar because they were applied to AIDS over a decade ago. Back then, these same four misconceptions were put forward as justifications for inaction. Fortunately they were successfully eradicated and expanded access to prevention and care for HIV/AIDS is now considered one of the greatest achievements in the history of global health. The same success can now apply to NCDs if we develop the right evidence-based policies and if we continue to involve all relevant actors.

NCDs are the driving force behind a health picture that can be characterized by two words: change and complexity. Our common challenge is that most health systems simply have not kept up with the pressures derived from the epidemiologic transition. In particular, ministers of health throughout the world are facing unprecedented demands as they seek to become effective stewards to develop health systems that respond to the needs and expectations of the population with equity, quality, and financial protection for all.

This complexity can only be addressed through a comprehensive response to NCDs built on three major pillars:

- First, the design and application of a new generation of health promotion and disease prevention strategies;
- Second, the achievement of universal social protection guaranteeing access to high-quality care without fear of financial catastrophe;
- Third, the adoption of innovations in the delivery of health services that make use of the technological and managerial revolutions of our times.

Many countries have made progress along these pillars. Mexico is a relevant example. In the following paragraphs, some of the most important lessons in the use of each of the pillars in a reform recently implemented in this country are discussed.

The first pillar was predicated on the notion that health systems will not be able to handle the growing burden of NCDs without a renewed emphasis on public health. Aware of this reality, a crucial component of the Mexican reform was the establishment of a new public health agency charged with protection of the population against health risks through food safety, definition of environmental and occupational standards, regulation of the pharmaceutical industry, and control of hazardous substances like alcohol and tobacco.

Along with other developments, this new agency has greatly strengthened the stewardship role of the Ministry of Health, which has become empowered to mobilize all instruments of public policy in the pursuit of health as a social objective.

In addition, the financial re-engineering of the health system included a protected fund for community health services targeting health promotion and disease prevention interventions, including, of course, those targeted at NCDs.

Important as promotion and prevention are, control efforts must also include access to health care. Indeed, even if we invest increasing amounts of resources in the prevention of NCDs, we will still need to deal with the consequences of exposures to risks that have already occurred. Those consequences include episodes of disease that require treatment, which all too often exposes families to the associated risk of financial catastrophe. For this reason, a comprehensive strategy must also include the second pillar: universal social protection.

Based on sound evidence about the extent of pernicious out-of-pocket payments, in 2003 the Mexican Congress approved a major legislative reform establishing a system of social protection in health. This system has substantially increased public funding for health in order to provide universal health insurance, including the half of the population, 50 million persons, most of them poor, who had lacked protection until then.

The vast majority of these persons are now enrolled in a new public insurance scheme called *Seguro Popular*, which guarantees access to a comprehensive package of cost-effective services covering the prevention, early detection, diagnosis, treatment, and palliation of the major causes of ill health, including, of course, NCDs. The law stipulates that the package must be progressively expanded and updated annually on the basis of changes in the epidemiologic profile, technological developments and resource availability.

The key to expand such resources has been to start with an explicit set of guaranteed benefits. This ties the reform to concrete deliverables, which is a main ingredient to gain public support. This approach tackles health system strengthening starting with the desired

outcomes, rather than with the existing inputs, as is the usual practice. Once the package of guaranteed interventions has been defined, it is possible to work our way backwards to estimate the requirements for inputs, including financial resources, workforce development, facilities, drugs, and other technologies.

Thanks to this approach, there was ample support for increasing public investments in health, despite general economic difficulties. The recipe for success was very simple: the Ministry of Health didn't *ask* for money; rather it *offered* explicit benefits for all, including the health benefits but also the large economic benefits of reducing the burden of chronic diseases.

An explicit package of interventions is the key to develop a "diagonal" strategy, whereby specific disease priorities are used to strengthen the overall structure and function of the health system.[21]

The true test of a reform, however, comes when benefits and resources make their way to the communities and facilities where actual delivery of services takes place. And this leads to the third and final pillar of health system strengthening: the deployment of innovations to assure that high-quality services reach all who need them. A particularly promising avenue is offered by the mobile phone revolution, with its enormous potential to expand access. Equally important are managerial innovations to improve efficiency, such as the delivery of NCD care in settings that require less intensity in the use of human resources and medical technology but still achieve good levels of quality.

4. Conclusions

In order to address the challenge of NCDs in the developing world we need to put in place a comprehensive strategy whose components have to be implemented both at the global and the local levels.

First of all, we need to overcome the lack of attention to this development challenge and integrate NCDs with communicable diseases in the global health agenda. The main objective in this regard should be to expand the MDGs to include health targets related to NCDs common in low- and middle-income countries, such as hypertension, diabetes, and cancer. WHO, in fact, has already proposed a 25% reduction of deaths attributed to NCDs by 2025 based on 2010 rates.

Second, it is necessary to mobilize local and global resources to finance the sustainable implementation of comprehensive strategies to address NCDs. Additional global resources will be crucial to implement these strategies in low-income countries.

Third, new health initiatives should consider the integration of prevention and treatment to control NCDs in a mutually reinforcing way. There are lessons to be learned in this respect from AIDS, where treatment has enormous impacts in preventing dissemination. Early detection and treatment of diabetes is also crucial to avoid the complications of this ailment, which require complex and costly interventions that impose pressures both on households and health systems. In reality there is no choice but to strengthen health systems so that they can offer comprehensive responses to the double burden of disease confronted by low- and middle-income countries.

The attention to the full cycle of care also implies the integration of all sectors whose activities are related to health in order to design and implement not only health policies but

also *healthy* policies. This integration is particular relevant to the control of NCDs since many of the risk factors related to them fall beyond the limits of the health sector.

Finally, the health community should strive to create networks that guarantee the continuity of care, which is a crucial component of the treatment of most chronic diseases. A related transformation involves moving beyond health *centers*, which by definition concentrate human and technological resources, into health *spaces*, which extend the reach of comprehensive care into schools, workplaces, recreational areas, and the homes of those who live with a chronic condition.

We should recognize that the driving force to face the NCD challenge will be located in countries. However, no individual nation can respond on its own to the global challenges that underlie the risk factors for NCDs. To address them we require international collective action in the pursuit of shared goals. A major vehicle in this respect is the development of global policy instruments, like the Framework Convention on Tobacco Control. Another crucial element comes in the form of global public goods, like the evidence base that must be built by rigorously evaluating national innovations. In this way, it will be possible to fuel a process of shared learning among countries about what works and in which context.

International action also requires the mobilization of global solidarity, as the fight against HIV/AIDS has so successfully exemplified. NCDs once again offer the world the chance to demonstrate that we are all committed to the universal value of health. Everyone has a role in this common endeavor: national governments, bilateral agencies, international organizations, global partnerships, private business, and the rich diversity of civil society, from professional associations and advocacy groups to academic institutions and research centers.

Giving NCDs their rightful place in the global health agenda will not be an easy task. Yet, if we act together to develop a prompt and comprehensive response, major improvements will be made in the health and wellbeing of the world population.

5. References

[1] Abegunde DO, Mathers CD, Adam T, Ortegon M, Strong K. The burden and costs of chronic diseases in low-income and middle-income countries. Lancet 2007;370:1929-38.

[2] WHO. Global measles deaths drop by 74%. Available at: http://www.who.int/mediacentre/news/releases/2008/pr47/en/index.html. Accessed June 10, 2011.

[3] Centers for Disease Control and Prevention. World Malaria Day 2011. Achieving progress and impact. Available at: http://www.cdc.gov/Features/WorldMalariaDay/. Accessed June 10, 2011

[4] IHME. Maternal mortality for 181 countries, 1980-2008: a systematic analysis of progress towards MDG5. Available at: http://www.healthmetricsandevaluation.org/research/publication-summary/maternal-mortality-181-countries-1980-2008-systematic-analysis-progress. Accessed June 10, 2011.

[5] UNDP. Human Development Report 2010. The real wealth of nations. Pathways to human development. New York: UNDP, 2010:143-47.

[6] World Health Organization. Global Status Report on Noncommunicable Diseases 2010. Geneva: WHO, 2011:1.
[7] Economist Intelligence Unit. Breakaway: The global burden of cancer- challenges and opportunities. London: The Economist Intelligence Unit, 2009.
[8] García M, Jemal A, Ward EM, et al. Global cancer facts and figures 2007. Atlanta, GA: American Cancer Society, 2007.
[9] Lozano R, Gómez-Dantés H, Lewis S, Torres-Sánchez L, López-Carrillo L. Tendencias del cáncer de mama en América Latina y el Caribe. Salud Pública de México 2009;51(supplement 2):S147-S156.
[10] World Health Organization. Cancer. Available at: http://www.who.int/mediacentre/factsheets/fs297/en/index.html. Accessed September 6, 2011.
[11] Goodarz D, Finucane MM, Lu Y, Singh G, Cowan MJ, Pachiorek CJ et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet 2011;378(9785):31-40.
[12] International Diabetes Federation. The economic impacts of diabetes. Available at: http://www.idf.org/diabetesatlas/economic-impacts-diabetes. Accessed September 6, 2011.
[13] Withrow and colleagues quoted in Wang YC, McPherson K, Marsh T, Gortmaker S, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. Lancet 2011;378:815-25.
[14] Swinburn BA, Sacks, G, Hall KD, McPherson KI, Finegood DT, Moodie M, Gortmaker SL. The global obesity pandemic: shaped by global drivers and local environments. Lancet 2011;378:804-14.
[15] Institute of Medicine. Early Childhood Obesity Prevention Policies. Washington, DC: IOM, 2011.
[16] Instituto Nacional de Medicina. Encuesta Nacional de Salud y Nutrición 2006. Cuernavaca, Morelos: Instituto Nacional de Salud Pública, 2006.
[17] King D. The future challenge of obesity. Lancet 2011;378:743-44.
[18] World Health Organization. About two out of five cancers can be prevented. Available at: http://www.who.int/mediacentre/multimedia/podcasts/2010/cancer_day_20100204/en/index.html. Accessed February 25, 2011.
[19] Brundtland GH. International Policy Conference on Children and Tobacco. Available at: http://www.who.int/director-general/speeches/1999/english/19990318_international_policy_conference.html. Accessed Febraury 27, 2011.
[20] Porter M. A strategy for health care reform. Toward a value-based system. New England Journal of Medicine 2009; 361:109-112.
[21] Sepúlveda J. Foreword. In: Jamison DT, Breman JG, Measham AR, et al, editors. Disease control priorities in developing countries [2nd edition]. New York: Oxford University Press for The World Bank, 2006: xiii-xv.

16

Diseases of Poverty: The Science of the Neglected

Pascale Allotey, Daniel D. Reidpath and Shajahan Yasin
Global Public Health, School of Medicine and Health Sciences, Monash University, Sunway Campus, Malaysia

1. Introduction

Diseases of poverty are those diseases identified as affecting the poorest and most disadvantaged populations in the world. Poverty is one of the main risk factors for the conditions, creating exposure to poor water and sanitation; poor nutrition, poor environmental conditions that favour the growth and spread of micro-organisms and vectors that cause and transmit disease; and lack of education and access to appropriate disease prevention, health promotion, treatment and rehabilitative services. Diseases of poverty include for instance, the neglected tropical (communicable) diseases (NTDs) which until relatively recently were considered low priority for both governments and pharmaceutical companies (1–4). Furthermore, diseases of poverty increasingly include the non-communicable diseases (5–7); hypertension, cardiovascular diseases, diabetes and other metabolic diseases and cancers, previously considered diseases of affluence (8–11). While there is some variation in the specific drivers that cause and exacerbate the communicable and non-communicable diseases for the poor, invariably, the processes and context are similar, impeding choices for healthier lifestyles, access to and acceptability and affordability of regular and quality care for chronic conditions and strategies for prevention and health promotion. In turn, affliction with these diseases hinders economic opportunities and development and perpetuates poverty. The disease increases vulnerability and exposure to poverty by increasing household expenditure and decreasing household income.

Through mechanisms provided by the Millennium Declaration and associated Millennium Development Goals, the World Economic Forum, the Global Fund, the Bill and Melinda Gates Foundation and the US President's Emergency Fund for AIDS Relief, the global health community has highlighted the plight of the poor and vulnerable, and gained support to address the major diseases. There is more funding available in global health now than there has ever been before (12–14). Major drug companies have committed to free donation of particular pharmaceuticals in an effort to achieve elimination of a number of diseases (15). The more recent UN Summit on NCDs employed this global advocacy process to elicit support from the highest levels of government to address the growing burden of specific chronic diseases. Critically however, programmes that result from these global health campaigns have historically been characterised largely by disease focused, vertical interventions that treat communities as a collective, providing a large scale clinical intervention. Much less attention is

focused on the more persistent underlying contributors to diseases of poverty – poverty and its other contextual drivers that are intimately interlinked with the diseases and outcomes.

In this discussion paper, we argue that despite the importance of these contextual drivers, they are largely neglected in the science and evidence that contributes to solutions for addressing diseases of poverty. We begin with the premise that there are fundamental differences in the ways that different disciplines conceptualise health, illness and disease. From a biomedical and clinical sciences perspective, diseases of poverty represent 'slugs, bugs and drugs' and present an ideal opportunity for technical fixes. There is robust evidence on the efficacy of these fixes and a strategy based on this evidence presents good value for money(16–20). From the perspective of the social sciences however, there is less of a separation between the person, the human condition, the environment and the disease process. The interest, from a social science perspective is primarily in the social, cultural, environmental and economic drivers of poverty and disadvantage, societal norms that mitigate marginalisation and the ecological factors that determine who becomes ill, what they do about it, and the outcomes of the illness. This would therefore also encompass the contextual factors that would enhance or hinder the delivery of a given biomedical strategy that involves populations. While robust and theoretically grounded, evidence from social science research and solutions that arise from that research may not necessarily present the kinds of context free, quantifiable, linear solutions that are frequently desired under biomedical research models. Similarly, under social science models, a solution that removes proximal causes of suffering without addressing the more distal and complex contextual factors that continue to put populations at risk, may not appear to be a desirable end point for a strategy. In this paper therefore we explore:

1. The contextual factors the define diseases of poverty;
2. The challenges in conceptualising and operationalizing these factors for the purposes of generating evidence;
3. The barriers to the translation of social science generated evidence in global public health; and
4. Some solutions to rebalancing the scientific approaches to neglected.

To address these questions, this report consists of a critical review of the diseases of poverty with a focus on the social, cultural, environmental and other contextual factors that affect risk, exposure, treatment and sequelae. In this context, diseases of poverty refer to the neglected diseases defined as those diseases which (i) have a disproportionate effect on the most disadvantaged sections of the community (the poor and marginalised); and (ii) lack investment in research and development for solutions that are explicitly accessible to the disadvantaged. We then provide a critical analysis of the sciences required to explore the complex nature of neglect in diseases of poverty and offer some suggestions for a broader approach to achieving long term solutions.

2. The context of diseases of poverty

Most of the conditions identified as diseases of poverty are treatable with currently available drugs. That notwithstanding, prevalence of these conditions remains high and the conditions persist (21). The neglected tropical diseases campaign for instance has relentlessly highlighted the plight of the populations affected by the range of target diseases.

A great deal has been made of stigmatization, disfigurement, persistent poverty, poor maternal and child health outcomes, poor health and education of children caused by infectious diseases (4,22–24). The choice of the word "neglect" is pointed and loaded, forcing us to reflect on our social obligations. Inherent in this campaign strategy is an appeal for the recognition of human suffering and the need for social justice (25).

These issues have been raised time and again by researchers working across the areas of health and human rights, the social determinants of health (26) anthropology and sociology (27–31) to mention a few. At the very least increasing standards of living, provision of the basic human rights of food, shelter, and clothing are definitive interventions towards the elimination of diseases of poverty. The body of evidence that supports the need for structural intervention is significant (32) and is obvious in the lack of these diseases in communities with an even marginally higher socio-economic status than "the bottom billion" (33). Tackling structural problems is harder because the interventions required are more complex; some have suggested too complex to consider (34). However not intervening at these levels increases the futility of current efforts. The re-emergence of diseases that were supposed to have been eradicated 40 years ago (35) is a case in point.

Other vulnerabilities highlighted in diseases of poverty include stigmatisation, social isolation, and disfigurement. These are vulnerabilities that result from social and cultural norms of what is considered normal and who is an acceptable member of the community (28,36,37). The effects of these on health relate to values that are less tangible than disease; equity, opportunity, access - and require intervention at different levels.

The basic concern here is not new and to a significant degree, revisits the major, largely unresolved debates that raged almost 40 years ago between proponents and opponents of Primary Health Care (PHC) (38–40). The critical question is this: does one partition out individual, proximal, biological causes (i.e., the disease) and address them as independent context free problems, or is there a need for a different approach which attempts to address the multiple distal and proximal causes within the context in which they occur? The primary health care debates addressed this question in favour of a disease specific approach with the introduction of Selective Primary Health Care programmes (41), vertical programmes. This establishes the putative 'pro-poor' credentials of diseases of poverty, despite the focus on identifying unabashed medical and technological fixes – the "magical bullet" to combat disease (38).

The contribution of the biomedical technologies cannot be underestimated. However, unless there are also significant interventions to address *health* and *poverty*, and the myriad marginalising factors in the social, cultural, economic, political and physical environments in which affected populations live, there will continue to be neglected people. Even in the research into the NTDs there is a distinct and patent disinterest in the social and contextual (42). Vaccines and drugs do not cure *neglect or poverty* and are not sufficient to rescue the neglected bottom billion from poverty (18).

3. The Implementation gap[1]

Even if it is decided that it would be safe to focus on the health side of the agenda rather than the poverty side, social and environmental (i.e., contextual) concerns cannot be

[1] This section draw significantly on earlier work of the authors and re-presents a number of the ideas without repeated citation, but also extends on some of those ideas(43).

avoided. An almost exclusive focus on the biomedical overestimates the value of the current science, leaving unresolved issues with implementation; that is, embedding a putatively effective intervention in a community. It is, after all, not enough to have the perfect cure if no one in need is able to receive it. Whether an intervention to be implemented in neglected populations has the same benefit in that population as it does in another population is an empirical question.

The randomised control trial (RCT) is widely regarded as the "gold standard" form of scientific evidence for establishing the effectiveness of a treatment (i.e., the cause effect relationship between treatment and cure), with decreasing levels of evidence treated with increasing levels suspicion. The problem with the RCT (and the levels of evidence) is that, in a general sense, and contrary to the expectations of many researchers, an RCT does not show the effectiveness of a treatment. It shows the effectiveness of a treatment in a particular context. Conducting multi-site RCTs, or conducting meta-analyses of multiple RCTs supports the generality of the finding. However, the conclusions about effectiveness can never be made without acknowledging the very controlled nature of experimental studies on which the conclusions about effectiveness are based; and by extension, the limitations imposed on generalising the results into less controlled, more realistic, contexts.

The intention to treat (ITT) analysis of RCTs is a partial acknowledgement of the problems of context. In the simplest kind of RCT, patients are randomly allocated to a treatment or a control (non-treatment / "usual treatment") group. Imagine that some people who were allocated to the treatment group ended up receiving no treatment – just like the control group. Under the ITT analysis, one analyses the results of the intervention as if all the people allocated to the treatment group, even those who did not receive treatment, did end up receiving treatment. This can seem somewhat counter-intuitive. Why would one analyse data counter to the reality of what happened? The analysis, however, establishes the effectiveness of a policy, i.e., an intention to treat patients in a particular way. The biological efficacy of the treatment should have already been established in early stage trials, and not be in doubt. The ITT analysis established the effectiveness of a treatment policy in a particular clinical setting.[2]

The use of community-based trials, and 'less rigorous' forms of effectiveness study try to capture the likely context in which an intervention might actually be employed; and to some degree they support the generalisation of the findings. A caveat, however, always remains, because study sites are inevitably different from sites that do not fall under the scrutiny of researchers. The context of the research study is not the context in which most lives are lived. The generalisation of the conclusions from the research study site to the populations that do not live under those conditions goes beyond the science.

The philosopher of science Nancy Cartwright raised points relevant to this argument in other branches of science. The issue is about what one knows in a general sense from doing scientific research. One of her points was that what one knows, relates to the context in which the research was conducted. Two illustrative examples of hers relate to the electronic transistor and to a leaf blowing in an alley. Consider the first example of the electronic transistor; a device used to regulate the flow of electricity. The basis of the transistor is grounded in quantum physics – a

[2] Interestingly, DDR recently read a description of statistical techniques to avoid the ITT analysis, so that the "true" effect of the intervention could be estimated. This presupposes that the idea of a true effect devoid of a context in which a treatment is applied makes any sense – which seems very doubtful.

theory that is free of contextual considerations. This means that a transistor works the same in New York, Bogotá, and Ouagadougou. When you start your laptop computer, which has millions of transistors, you do not first have to find out where to make contextual adjustments to the transistors. Superficially the science under-pinning the transistor looks to provide the very kinds of context-free insight that real science is all about.

On reflection the context-free nature of the findings are superficial. It is not the case that the transistor works in all contexts; rather, industrial manufacturing processes have been developed which make sure that the context within the transistors' housings remain the same without regard to where the transistors are. In effect, manufacturers have learned to create miniature, identical, controlled environments, with a fixed context of operation that conforms to an idealised model. The quantum effects work reliably and consistently within the bounds of the miniature environment, but without the same certainty outside that environment.

The second example is of a leaf. Science and engineering has provided significant insights into aerodynamics. We have instrumentally valuable theories that predict airflow and lift. Empirical work in wind tunnels, computer simulation efforts and theoretical advances allow for very precise predictions to be made about how aircraft will behave under a range of plausible environmental conditions. Predicting the path, however, that a leaf will follow when blown down an alley is beyond us. The idealised understanding that we have of aerodynamics allows us to frame and control the context of the science that is done. Aircraft wings are crafted so that they maximise our predictive capacity, and conform to our understandings of the laws of aerodynamics. When we cannot control the context of the science, however, what we actually know becomes far less impressive.

These observations are not pedantry, and they do not belittle the science that allows us to fly aircraft and build computers. What they do suggest, however, is that our science works because we know and understand the context in which it is applied. With a change in context, the success of the science is less certain. When developing health interventions, we do not have the luxury of constructing the context to suit the kinds of interventions or designing the intervention to work in a single context. Rather, we need to engage in the type of science that embraces interventions that are contextually appropriate.

At a recent scientific meeting on community directed ivermectin distribution program for the control of Onchocerciasis, a report was presented from Nigeria where the intervention was not achieving the results anticipated given known effectiveness and the reported high coverage of ivermectin. When the gap between coverage and results was investigated, the evaluation team found that the villagers were receiving the ivermectin; however, instead of taking the tablets themselves, they were distributing them among their cattle. The villagers had decided that the economic benefit of a healthy herd far out-weighted the health loss they faced by failing to treat their personal affliction with onchocerciasis.

The science had shown that ivermectin was a clinically effective approach to onchocerciasis control in one context. Community-based trials confirmed the effectiveness after scaling up the intervention in another context (44); and the economic analysis showed that it was cost-effective (45,46). This was the 'truth' as revealed by the science of fixed contexts. The reality, however, was that the effectiveness of the intervention depended on a range of contextual factors – such as competing economic incentives. Having located the research in fixed (or well regulated) contexts, the likely variability of outcome that occurs in the wilds of real life, did not enter into any decisions about effectiveness.

There are two important corollaries to this. The first: imagine two interventions both of which are significantly more effective than no treatment. Furthermore, in clinical trials researchers have established that intervention A is significantly more effective than intervention B (i.e., $A > B > 0$). When the context changes from the controlled research environment to point of implementation, the apparent magnitude of the effect of the interventions can reverse, with intervention B having a greater effect than intervention A (i.e., $B > A > 0$ or $B > A = 0$). This will occur if, at the stage of implementation, the more effective A cannot be embedded in the community.

The second corollary, which is an extension of the first, is that interventions that seem to be cost-ineffective in one context maybe the cost-effective interventions in another context, and the cost-effective intervention in another context will be the cost-ineffective intervention in this context. Continuing to use interventions A and B, following the effectiveness studies, the economic analysis established that A is more cost-effective than B. However, on implementation, when A fails to achieve any community up-take, B becomes the more cost-effective of the interventions. The implications of this are hard to under-estimate.

Decision making based on effectiveness and cost-effectiveness, which is a rational approach to the optimal allocation of scarce resources, may fail dramatically if the information on which the decision is based comes from the partial science of fixed contexts.

As Allotey et al. observed (p.3), effectiveness is regarded as the appropriate end point for most intervention research. But knowing that a treatment is effective in routine clinical care is not enough, particularly in resource poor settings (i.e., the settings of the neglected). The goal must be the sustainable adoption of the intervention by the health systems and the target population, and not simply the establishment of effectiveness in a monitored clinical population. In other words, an intervention must become embedded; firmly integrated as part of the health system and the health culture of the disease endemic setting. It must be available, acceptable, accessible and affordable to those who need it; used appropriately, and become a part of the disease prevention, treatment seeking culture.

Biomedical research is neither intended to address nor capable of addressing questions about implementation. Thus, not only is the value of the biomedical research limited by our lack of research on the contextual effects associated with implementation, it is also outside the expertise of those scientists to address the issues.

4. The science of the neglected

To this point we have argued that the approach to the neglected diseases has leveraged the idea of the vulnerable and neglected population to advance an argument for providing additional resources to the biomedical scientists so that they can develop cures for neglected diseases – "vaccines against poverty". We then discuss the evidence about social vulnerability to disease, and the possibility of social interventions that address more distal causes of disease – intervening before the biomedical concerns arise. Finally we argued that the focus on proximal interventions is based on a flawed notion of the under-lying science and the generality of that science. In effect we argue for the development of contextually relevant science capable of accounting for social and environmental factors affecting the implementation of interventions.

What is missing from our discussion is (i) the research that supports the implementation of proximal cures, and (ii) the research that supports distal interventions that change the social

vulnerability of neglected populations to disease. The obvious place to look for this research is in the social sciences literature, or the intersection between the clinical, biomedical, and social sciences literature.

In a bibliometric analysis of four diseases of poverty (chikungunya, dengue, leishmaniasis, and onchocerciasis) we found that social sciences contribute to less than 2% of the published research (42). That was a generous counting of the social sciences contribution. The research that was funded was generally insipid, because it was there to act as a hand-maiden for biomedical research, never intended to support a research agenda of implementation or distal intervention. And the lack of a social sciences research agenda has a negative impact on the value of the biomedical research that is conducted, and limits our options for intervention to proximal cures.

To say that the social sciences have been totally overlooked in the global health efforts would however be inaccurate. The value of the social sciences up until now, however, is qualified. In the area of NTDs, evidence from anthropological studies on stigmatization, the lived experiences of patients disfigured by diseases such as leprosy, yaws, onchocerciasis and filariasis, and the effects of these on health seeking, access to and quality of care, have been used particularly to support advocacy (4,18,28,53,24,54,55). The research that explores the reasons for the failures of programmes for instance is not insubstantial. Anthropological research has provided data on the importance of cultural and social constructions of illness and disease. We have some understanding of the different levels of practitioners, how and why they might be consulted and their role (or lack thereof) within a formal health system. There is evidence from the social sciences of the complexities and pathways to health seeking, the economic and social drivers, the effects of gender and other social determinants. Health economics has shed light on willingness of patients or clients to pay for different types of health services, interventions and pharmaceuticals; and the local market forces that hinder or enable distribution of and access to health services and pharmaceuticals. Health services and health systems research provides rigorous data on the socio-economic and political context in which local, national and global health policy supports (or otherwise) disease control programs.

In broad terms however, social science research in this area has to date focused largely on the evaluation of the implementation process and on factors that will enhance community participation in community based programs (56). Both the process and the outcome indicators therefore relate to the administration of treatment and where appropriate, a short term reduction in NCDs. In other words these approaches to 'deploying' the social sciences are rather utilitarian and often tokenistic (43). The consequences to this are the often questionable quality of the social science evidence generated. Implementation research for instance, if well designed and implemented has the potential to contribute significantly to disease control efforts – however it is an area of research that is poorly funded (43) The problem arises often because social scientists are invited onto teams to undertake specific research projects rather than being a conceptual part of the planning of the intervention (27)

To obtain the higher objective of improving the health and reducing vulnerabilities, it is important for researchers, policy makers and funding agencies to broaden the perspective on the range of research that is needed to address neglected diseases of neglected populations, and to rethink the types of integrated interventions and the nature of evidence to show effectiveness. There is a need to refocus on the health of neglected populations - health as an enabling process (38) - and not merely removing disease.

Critical opportunities are missed through the lack of integration of data from the social science disciplines. Health and illness are social constructs and as such, the disciplines and theories that help us to make sense of these issues should be as much a part of the agenda as pharmaceutical developments. It is tragic, for instance that so much is made of the suffering of patients of neglected tropical diseases, but there is little if any evidence in the funded programmes that addresses how families and communities affected by these diseases could be supported to deal with the social and economic sequelae. Studies of outbreaks of infectious diseases in South East Asia also highlight the almost exclusive disease focus of public health interventions and the total neglect of the mental health and social and economic consequences of these interventions (described as social chaos) on the populations affected (57). To address these issues would require a more complex understanding of the community and its dynamics and the broader political context in which the affected populations live.

Studies in gender for instance have produced frameworks that facilitate the integration of gender across programmes. Similar approaches have been suggested for use with the social sciences (27,43,57,58)

5. Alternative models

There are essentially two issues that are conflated in the advocacy and the current approach to diseases of poverty. The first is the focus on neglect and vulnerabilities – as highlighted above, a significantly complex issue which we, as global health professionals, have an obligation to address (47). These issues cannot however, be fully addressed by vertical programmes. The second is the specific issue of disease which forms an important part of the factors which may be the cause of, but also exacerbate and sustain poverty and vulnerability. This issue is the focus of vertical programmes (41). Interventions to address these two issues should clearly not be mutually exclusive, but often are.

The question of which general approach is better does depend on the expected outcomes but may of course be empirical. Assuming that the expected goal, as most global health programs stipulate, is the improvement of the health of populations, how would a poverty reduction, empowerment, equity based development programme fare against a preventive chemotherapy programme for instance, or one that combined approaches. Studies that test this empirically are rarely designed, in part because the different interventions seek different outcomes. Vertical programmes measure success in terms of reductions in the occurrence of specific diseases. Contextually based, comprehensive programmes count some broader measure of well-being as the desirable outcome. However it is difficult to imagine that there would be no value added to ensuring that the pieces lock together seamlessly. Programmes that privilege longer term improvements in the living conditions over merely achieving significant coverage of mass drug administration have shown a greater impact in rescuing communities and tackling concerns about neglected diseases and neglected populations (48). These tend to be smaller programmes, with significant input from communities and do not operate under the pressures of reporting to funders. Furthermore, when the outcomes of such programmes are published, the robustness of the 'evidence' is often questioned because they were not designed as 'empirical' studies (4,49–51).

There are data that could arguably have the potential to provide a proxy indication of how the different approaches measure up. We know for instance that significant funds have been invested into global public health most of which have gone into vertical programmes dealing with the big three and more recently, the neglected tropical diseases (13,52). Data

are also available on investments into other programmes designed to meet the other millennium development goals, which also address the vulnerabilities highlighted by the neglected disease advocates. A cost effectiveness analysis of these investments could technically provide an indication of what a dollar could purchase per intervention type. However the success of programmes still tends to be measured often by their coverage rather than by longer term outcomes, and in global health, seldom by improvements in the levels of poverty and broader development. Reasons for this include the time limited nature of programmes; the discipline focus of people involved in programmes, that is health sector and therefore the disease focus – lack of capacity to design the relevant research, monitoring and evaluation tools that would allow a focus that were any broader.

To focus on the addressing neglect and vulnerabilities from a health perspective would require a different way of conceptualising the link between poverty, health and disease, acknowledging the complexities and developing appropriate and realistic solutions. This would mean more than a simple combination of individual supplementary (vertical) programmes. It would also necessarily require a redefinition of outcomes and successes, working to a longer time frame than is currently adhered to in disease based vertical programmes. A detailed discussion is beyond the scope of this paper.

6. Conclusion

Diseases of poverty represent a rich and dynamic interplay between the context of people's lives and the disease process. The interaction is complex and evolves within a social and cultural context as much as it does within a physical and biological context. Understanding this complex dynamic is crucial for the sustainable management of diseases of poverty. The evidence from the health literature, however, is that there is little investigator driven social science research to speak of in the diseases of poverty, and a similarly poor presence of interdisciplinary science. Without this, our understanding and management of diseases of poverty is inevitably reduced to a strategy that relies on a repetitive, reductionist flat-world science to overcome an acknowledged complex system.

The research to address neglected diseases of poverty needs more sophisticated funders and priority setters. Pharmaceuticals (including vaccines) are critical, but they are not the only solutions, and their final application is not in flat worlds. Their application is in complex dynamic worlds in which pathologies evolve to exploits the social nature of humans. Our current understanding of the dynamic, and our understanding of how to develop sustainable approaches to disease management are poor. There are no research templates to overcome this, and the silos of current science into the diseases of poverty have discouraged the development of genuinely interdisciplinary research.

As a major recommendation there is a need to reconceptualise the outcomes for addressing vulnerability and the addressing the health needs of the neglected, poor, disenfranchised and dispossessed. Recognising that the challenges cannot be reduced to simplistic biomedical solutions is a first step. Global public health is ideally placed to bring together the different disciplines to engage in these developments.

7. References

[1] A New Era of Hope for the World's Most Neglected Diseases. PLoS Medicine. 2005 Sep 1;2(9):e323 EP -.

[2] Molyneux DH. Neglected tropical diseases--beyond the tipping point? The Lancet. 2010 Jan 2;375(9708):3–4.
[3] Hotez PJ, Kamath A. Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden. PLoS Negl Trop Dis. 2009;3(8):e412.
[4] Hotez PJ. Stigma: The Stealth Weapon of the NTD. PLoS Negl Trop Dis. 2008 Apr 30;2(4):e230.
[5] Lopez AD, Mathers CD. Measuring the global burden of disease and epidemiological transitions: 2002-2030. Ann Trop Med Parasitol. 2006 Sep;100(5-6):481–99.
[6] Das S. Rising trend of non-communicable diseases in low socioeconomic areas. Natl Med J India. 2007 Dec;20(6):319.
[7] Schneider M, Bradshaw D, Steyn K, Norman R, Laubscher R. Poverty and non-communicable diseases in South Africa. Scand J Public Health. 2009 Mar;37(2):176–86.
[8] Gwatkin DR, Guillot M, Heuveline P. The burden of disease among the global poor. Lancet. 1999;354(9178):586–9.
[9] de-Graft Aikins A, Unwin N, Agyemang C, Allotey P, Campbell C, Arhinful D. Tackling Africa's chronic disease burden: from the local to the global. Globalization and Health. 2010;6(1):5.
[10] World Health Organization. Noncommunicable diseases country profiles 2011 [Internet]. Available from: http://bit.ly/nG9Hu8
[11] The global burden of chronic diseases [Internet]. [cited 2011 Sep 19];Available from: http://www.who.int/nutrition/topics/2_background/en/index.html
[12] McCoy D, Kembhavi G, Patel J, Luintel A. The Bill & Melinda Gates Foundation's grant-making programme for global health. Lancet. 2009 May 9;373(9675):1645–53.
[13] Ravishankar N, Gubbins P, Cooley RJ, Leach-Kemon K, Michaud CM, Jamison DT, et al. Financing of global health: tracking development assistance for health from 1990 to 2007. Lancet. 2009 Jun 20;373(9681):2113–24.
[14] Crane BB, Dusenberry J. Power and Politics in International Funding for Reproductive Health: the US Global Gag Rule. Reproductive Health Matters. 2004 Nov;12(24):128–37.
[15] Alleviating The Suffering Caused By River Blindness And Lymphatic Filariasis (Elephantiasis) | Mectizan Donation Program [Internet]. [cited 2011 Sep 19];Available from: http://www.mectizan.org/
[16] Musgrove P, Hotez PJ. Turning Neglected Tropical Diseases Into Forgotten Maladies. Health Aff. 2009 Nov 1;28(6):1691–706.
[17] Hotez P, Bethony J, Brooker S, Albonico M. Eliminating neglected diseases in Africa. The Lancet. 365(9477):2089.
[18] Hotez PJ, Fenwick A, Savioli L, Molyneux DH. Rescuing the bottom billion through control of neglected tropical diseases. The Lancet. 373(9674):1570–5.
[19] Rosenfield PL, Golladay F, Davidson RK. The economics of parasitic diseases: Research priorities. Social Science & Medicine. 1984;19(10):1117–26.
[20] Hotez PJ, Molyneux DH, Fenwick A, Ottesen E, Ehrlich Sachs S, Sachs JD. Incorporating a Rapid-Impact Package for Neglected Tropical Diseases with Programs for HIV/AIDS, Tuberculosis, and Malaria. PLoS Medicine. 2006 May 1;3(5):e102 EP -.
[21] WHO | Innovative and Intensified Disease Management (IDM) [Internet]. [cited 2009 Dec 10]; Available from: http://www.who.int/neglected_diseases/disease_management/en/index.html

[22] Conteh L, Engels T, Molyneux DH. Socioeconomic aspects of neglected tropical diseases. The Lancet. 2010 Jan 16;375(9710):239–47.
[23] Hotez PJ. Empowering Women and Improving Female Reproductive Health through Control of Neglected Tropical Diseases. PLoS Negl Trop Dis. 2009 Nov 24;3(11):e559.
[24] Hotez P. Measuring Neglect. PLoS Negl Trop Dis. 2007 Nov 28;1(2):e118.
[25] Allotey P, Reidpath DD, Pokhrel S. Social sciences research in neglected tropical diseases 1: the ongoing neglect in the neglected tropical diseases. Health Res Policy Syst. 2010;8:32.
[26] Marmot M, Friel S, Bell R, Houweling T, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. The Lancet. 2008 Nov;372(9650):1661–9.
[27] Manderson L, Aagaard-Hansen J, Allotey P, Gyapong M, Sommerfeld J. Social research on neglected diseases of poverty: continuing and emerging themes. PLoS Negl Trop Dis. 2009;3(2):e332.
[28] Perera M, Whitehead M, Molyneux D, Weerasooriya M, Gunatilleke G. Neglected patients with a neglected disease? A qualitative study of lymphatic filariasis. PLoS Negl Trop Dis. 2007;1(2):e128.
[29] Dunn FL. Role of human behavior in control of parasitic diseases. Bulletin of the World Health Organization. 1979;57(6):887–902.
[30] Dunn FL. Behavioural aspects of the control of parasitic diseases. Bulletin of the World Health Organization. 1979;57(4):499–512.
[31] Manderson L, Jenkins J, Tanner M. Women and tropical diseases : Introduction. Social Science & Medicine. 1993 Aug;37(4):441–3.
[32] Lynch JW. Income inequality and mortality: importance to health of individual income, psychosocial environment, or material conditions. BMJ. 2000 Apr;320(7243):1200–4.
[33] Smith GD, Lynch J. Commentary: Social capital, social epidemiology and disease aetiology. Int. J. Epidemiol. 2004 Aug 1;33(4):691–700.
[34] Meyers W, Portaels F. Mycobacterium ulcerans infection (buruli ulcer). In: Guerrant RL, Walker D, Weller PF, editors. Tropical Infectious Diseases. Principles, Pathogens and Practice. Philadelphia: Churchill Livingstone Elsevier; 2006. p. 429–35.
[35] Asiedu K. Yaws eradication: past efforts and future perspectives. Bull World Health Organ. 2008 Jul;86(7):499–499.
[36] Reidpath DD, Chen K, Gifford S, Allotey P. He hath the french pox: stigma, social value and social exclusion. Sociology of Health and Illness. 2005;27(4):468–89.
[37] Yang LH, Kleinman A, Link BG, Phelan JC, Lee S, Good B. Culture and stigma: Adding moral experience to stigma theory. Social Science & Medicine. 2007 Apr;64(7):1524–35.
[38] Rifkin SB, Walt G. Why health improves: Defining the issues concerning [`]comprehensive primary health care' and [`]selective primary health care'. Social Science & Medicine. 1986;23(6):559–66.
[39] Newell KW. Selective primary health care: the counter revolution. Social Science & Medicine. 1988;26(9):903–6.
[40] Magnussen L, Ehiri J, Jolly P. Comprehensive Versus Selective Primary Health Care: Lessons For Global Health Policy. Health Aff. 2004 May 1;23(3):167–76.
[41] Walsh J, Warren K. Selective primary health care: an interim strategy for disease control in developing countries. N Engl J Med. 1979 Nov 1;301(18):967–74.

[42] Reidpath DD, Allotey P, Pokhrel S. Social sciences research in neglected tropical diseases 2: A bibliographic analysis. Health Res Policy Syst. 2011;9(1):1.

[43] Allotey P, Reidpath D, Ghalib H, Pagnoni F, Skelly W. Efficacious, effective, and embedded interventions: Implementation research in infectious disease control. BMC Public Health. 2008;8(1):343.

[44] The CDI Study Group. Community-directed interventions for priority health problems in Africa: results of a multicountry study. Bulletin of the World Health Organization. 2010 Jul 1;88:509–18.

[45] Community-directed interventions for integrated delivery of a health package against major health problems in rural Uganda: perceptions on the strategy and its effectiveness [Internet]. [cited 2011 Sep 12];Available from: http://www.internationalhealthjournal.com/article/S1876-3413(10)00053-7/abstract

[46] Atun R, de Jongh T, Secci F, Ohiri K, Adeyi O. A systematic review of the evidence on integration of targeted health interventions into health systems. Health Policy and Planning. 2010 Jan 1;25(1):1–14.

[47] Lowry C, Schuklenk U. Two Models in Global Health Ethics. Public Health Ethics. 2009 Nov 1;2(3):276–84.

[48] Partners In Health (PIH), Health Care for the Poor [Internet]. [cited 2009 Dec 18];Available from: http://www.pih.org/home.html

[49] Kuper H, Solomon AW, Buchan J, Zondervan M, Foster A, Mabey D. A critical review of the SAFE strategy for the prevention of blinding trachoma. Lancet Infect Dis. 2003 Jun;3(6):372–81.

[50] Krieger N. Questioning epidemiology: objectivity, advocacy, and socially responsible science [editorial; comment]. Am J Public Health. 1999;89(8):1151–3.

[51] Hotez PJ. Training the Next Generation of Global Health Scientists: A School of Appropriate Technology for Global Health. PLoS Negl Trop Dis. 2008;2(8):e279.

[52] McCoy D, Kembhavi G, Patel J, Luintel A. The Bill & Melinda Gates Foundation's grant-making programme for global health. Lancet. 2009 May 9;373(9675):1645–53.

[53] Hotez P, Ottesen E, Fenwick A, Molyneux D. The neglected tropical diseases: the ancient afflictions of stigma and poverty and the prospects for their control and elimination. Adv Exp Med Biol. 2006;582:23–33.

[54] PLoS Neglected Tropical Diseases: Holidays in the Sun and the Caribbean's Forgotten Burden of Neglected Tropical Diseases [Internet]. [cited 2008 Nov 17];Available from: http://www.plosntds.org/article/info%3Adoi%2F10.1371%2Fjournal.pntd.0000239

[55] Muela Ribera J, Peeters Grietens K, Toomer E, Hausmann-Muela S. A Word of Caution against the Stigma Trend in Neglected Tropical Disease Research and Control. PLoS Negl Trop Dis. 2009 Oct 27;3(10):e445.

[56] Parker M, Allen T, Hastings J. Resisting Control of Neglected Tropical Diseases: Dilemmas in the Mass Treatment of Schistosomiasis and Soil-Transmitted Helminths in North-West Uganda. Journal of Biosocial Science. 2007;40(02):161–81.

[57] Phua K-L, Lee LK. Meeting the Challenge of Epidemic Infectious Disease Outbreaks: An Agenda for Research. Journal of Public Health Policy. 2005;26(1):122–32.

[58] Phua K-L. Fighting the Battle Against Infectious Diseases: Contributions of Selected Social Science Disciplines. Infectious Diseases: Research and Treatment. 2009 Nov 9;2009(1728-IDRT-Fighting-the-Battle-Against-Infectious-Diseases:-Contributions-of-Sele.pdf):5.

17

Health-Longevity Medicine in the Global World

Dan Riga[1], Sorin Riga[1],
Daniela Motoc[2], Simona Geacăr[3] and Traian Ionescu[4]
[1]*Academy of Romanian Scientists, Dept. Stress Res. & Prophylaxis, Al. Obregia Clin. Hosp. Psychiatry, Bucharest,*
[2]*Centre Appl. Physiol. & Mol. Biol., V. Goldis Western University, Arad,*
[3]*Faculty of Psychology and Education Sciences, University of Bucharest, Bucharest,*
[4]*Romanian Academy of Medical Sciences, Bucharest, Romania*

The doctor of the future will give no medicine,
but will interest his patient in the care of the human frame,
in diet and the cause and prevention of disease.
Thomas Alva Edison (1847-1931),
American inventor, scientist, and businessman

1. Introduction

1.1 Health in ontogenesis

Health in the human life cycles produces healthy longevity. The construction of health-longevity can be accomplished through primary prophylaxis, namely education, promotion, training, protection and prevention.

As such, medicine seeks to achieve the prevention of disease; it aspires to treat all pathologies, as secondary prophylaxis and leads to recovery after illnesses, as tertiary prophylaxis.

The common elements of longevity health sciences - LHS (Cutler et al., 2005a) and mental health - MH (Knapp et al., 2007) consist of personal sanogenesis at an individual level, and public health in relation to the societal dimension.

1.2 Objectives for health-longevity medicine. Past, present and future

Nowadays it is time to promote and apply the ancient wisdom concerning health and healing concepts alongside medical ones.

Actual scientific data about health strategy (human biology and risk factors, behaviour and lifestyle, health care systems, the environment), technological medical progress, information

technology and information technology and communication (ITC), together with experience and the practice of developed countries should be integrated at a regional and global level.

In addition, new concepts can be used and applied to the understanding of the complexity of healthy-longevity medicine at a global level:

- regional and global programmes, strategies and actions (WHO, Regional Office for Europe, 1986; Knapp et al., 2007);
- new paradigms for medical education: from health promotion and protection to longevity health sciences and life extension (S. Riga et al., 2010d);
- the implementation of a healthy diet and a physically active lifestyle (Simopoulous, 2005);
- nutraceuticals and nutrigenomics (D. Riga and S. Riga, 2011b);
- the palaestra paradigm (D. Riga and S. Riga, 2010a);
- synergistic anti-stress, anti-impairment and anti-ageing drugs and strategies (D. Riga and S. Riga, 1995-2005);
- regenerative and pro-longevity medicine (de Grey, 2004; D. Riga et al., 2004a; S. Riga et al., 2004b);
- a new pyramid of health-longevity services (S. Riga et al., 2011a);
- health, longevity and ecology - an integrated paradigm (D. Riga et al., 2010b);
- the bio-psycho-socio-ecological dimension of human being (S. Riga et al., 2010c).

2. Health and preventative medicine in ancient times

2.1 Prophylaxis and physical activity in traditional Chinese medicine

Dating back thousands of years, the practice of traditional Chinese medicine includes Yinyangism and Daoism as philosophical concepts, holistic and integrative medical concepts, phytotherapy (herbal medicine) and dietary therapy, acupuncture, Shiatsu and Tui na massage, movement therapy, Qigong, Taiji and other methods of maintaining health and vitality.

A remarkable characteristic of the Chinese system of natural healthcare is its prophylactic side. A programmatic document in this direction is the first Chinese medical text (c2600 B.C.). It stipulates: *Superior doctors prevent the disease. Mediocre doctors treat the disease before evident. Inferior doctors treat the full-blown disease* (Unschuld, 2003). In addition, this famous manuscript *Huángdi Neijing Suwen (Inner Canon: Basic Questions),* also known as *The Inner Canon of Huángdi* or *Yellow Emperor's Inner Canon*, book written between 2698 B.C. - 2596 B.C. presents a dialogue between the Yellow Emperor (Huángdi) and Qibo (Qi Bo, Chi Bo), his minister and advisor, an excellent physician and the father of massage treatment. Another quote from this treatise shows the importance of prophylaxis: *To treat an illness after it has already set in or to smother a riot already spread is liked digging for a fountain when you're already thirsty or making weapons after the war has already begun. Isn't it too late, I wonder?* (Lin, 2000).

A further defining feature of the traditional Chinese therapeutic system is the promotion of movement and physical activity in maintaining health and treating illness. The famous Chinese physician Huà Tuó (c145 A.D. - c208 A.D.), the first person in China to use anaesthesia in surgery, created a series of exercises called *Wuqinxi or Frolics* (*Exercise*) *of the Five Animals*, towards the end of the 2nd Century A.D. The exercises mimicked the

movements of the tiger, the deer, the bear, the monkey and the crane. In Huà Tuó's medical system, the therapeutic use of movement was inspired from nature: *Running water never grows stale, and the doorpost is never eaten away by wood decay. For the same reason, if we do physical activity on a regular basis, we can remain in good health and keep disease away. Regular exercises stimulate blood flow and the circulation of the qi* (*energy*), *thus maintaining the agility of the body* (Lin, 2000).

2.2 Preventative medicine in Greek and Roman antiquity

The doctrines of Mediterranean ancient medicine are also based on dietary (rational nutrition) and physical exercises.

Hippocrates of Cos (Kos), c460 B.C. - c370 B.C., one of the most outstanding figures in the history of medicine, emphasized the importance of diet: *Let thy food be thy medicine and thy medicine be thy food* (Hanson, 2006). Moreover, the veneration of the human body as well as daily and professional physical activity were extensively spread in Hellas. Palaestra (special arranged places and also a type of physical exercises) and the Ancient Olympic Games are only two examples.

The Romans, who conquered, took over and enriched Greek civilization, also had great respect for a harmonious development of the human body. Besides this, they pointed out the necessity and simultaneity of the sanogenetic binomial psychic ↔ body. The old adage: *Mens sana in corpore sano, Satyrae X (Book IV, Satyrae X, Line 356 - 10.356),* Decimus Iunius Iuvenalis (c60 A.D. - c135 A.D.), Roman poet, in still famous and up-to-date even now (D. Riga et al., 2009c).

The principles of preventative medicine and competitive health-vitality have been well-documented in human history since ancient times. Unfortunately, current civilizations and human beings could not manage, up to the present, to transform these principles into their daily routine or integrate them into their lifestyles.

3. From health to disease

3.1 Stress bio-medicine

From this perspective, the strategic key in public health is represented by stress medicine (stressology), adaptology and MH (S. Riga and D. Riga, 2008).

Figure 1 shows the multi-factorial progress, which localizes stress bio-medicine at the boundary/interface between normality-health-longevity and ageing-disease.

The integrative concept (from molecule to individual and society) groups together:

- the diseases of lifestyle/adaptation/civilization (Selye, 1976);
- stress-related disorders, burnout and chronic fatigue syndrome (Cooper, 1996; WHO. ICD-10, 1992);
- the free radical theory of ageing and free radical diseases (Harman, 1984);
- The oxidative stress theory of ageing and oxidative stress-associated diseases (Cutler et al., 2005c);
- antioxidant deficit diseases (in food, blood and tissues) (Slater and Block, 1991; Muller et al., 1992).

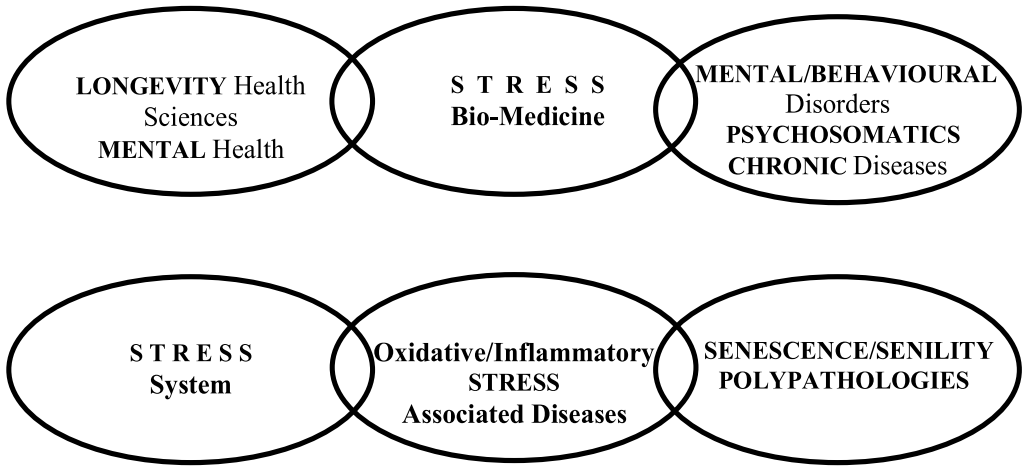


Fig. 1. Oxidative stress in stress bio-medicine, health, ageing and disease

In conclusion, the new concept of stress in bio-medicine represents the primary cause (the beginning) of various human illnesses: pathological manifestations of acute and chronic psychic stress, stress-related disorders, free radical diseases, oxidative stress-associated pathologies, accelerated impairment and ageing (premature senescence), diseases of lifestyle and civilization, nervous and body inflammatory-degenerative pathologies and senility.

3.2 Antagonism of health construction versus human pathology

Public health strategies and policies, as well as everyday preventative-prophylactic and medical-curative practice, are substantiated in dynamics by two opposite tetrads (cascades), (D. Riga and S. Riga, 2007; S. Riga et al., 2009a). These concepts also represent two antagonistic fundamental pathways:

- *stress* ↔ *ageing,* entropic, aetio-pathogenic *tetrad:* distress/stress-dependent disorders ↔ wear and tear/impairment ↔ premature/accelerated ageing ↔ poly-pathology;
- *health* ↔ *longevity,* anti-entropic, protective-therapeutic *tetrad*: anti-stress/eustress/adaptation ↔ anti-impairment/vitality/resistance ↔ anti-ageing/active, healthy longevity ↔ anti-illnesses/anti-diseases.

Therefore, health construction is in total opposition to the development of human pathology. Health construction promotes and protects sanogenesis and impedes the appearance and evolution of disease.

3.3 Dynamic structure of destructive cascade

Stress ↔ *ageing tetrad (distress* ↔ *impairment* ↔ *ageing* ↔ *disease)* is a progressively destructive, entropic and time-dependent phenomenon: from primary processes and chronic manifestations (distress, impairment, ageing) to chronic illnesses. The dynamic pattern of this cascade is shown in **Figure 2** (D. Riga and S. Riga, 2007).

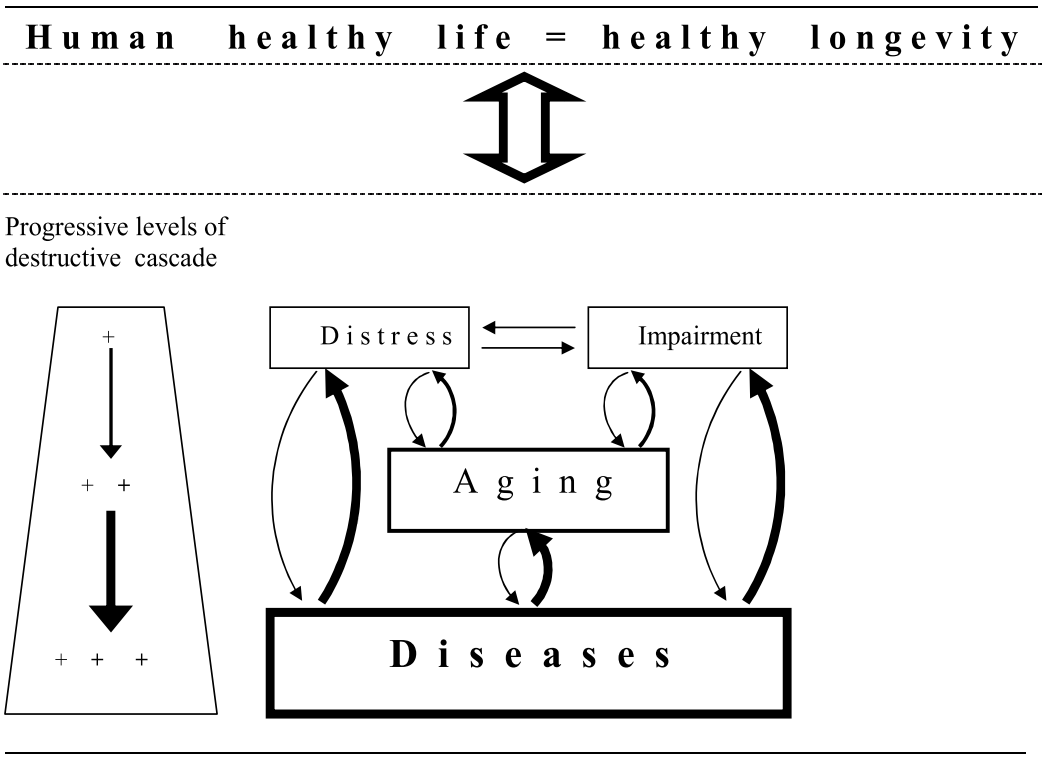


Fig. 2. Dynamic structure of destructive cascade:
distress ↔ impairment ↔ ageing ↔ disease.
From human healthy life/longevity to old age/poly-pathologies

Time acts in a very complex way:

- as a harmful amplifier – the initial subclinical stages turn into final clinical phases, namely into manifest diseases; and
- as a continuous initiator, by transforming causes into effects, which in their turn become secondary and multiple causes for new negative consequences; thus, the four components of the cascade successively represent both cause and effect.

In addition, free radical attacks, oxidative stress and antioxidant deficits are amplifiable and worsen in accordance with a pattern of destructive synergism. Therefore, the accumulation of distress, impairment and ageing is aggravated in oxidative stress (chronic) diseases (Cutler, 1996; Miwa et al., 2008).

3.4 Risk factors and preclinical stages of ageing and disease

"Risk factor" (an epidemiological concept) is a variable (characteristic, condition or behaviour) associated with an increased risk of disease (or infection, or injury). Sometimes, "determinant" is also used, being a variable associated with either increased or decreased risk. Risks factors are co-relational and not necessarily causal, since correlation does not imply causation.

They are categorized into *intrinsic* "within oneself" and *extrinsic* "outside" influences.

In another classification, risk factors are divided into four domains:

- *biological risk factors.* Firstly, they are represented by *age, gender and race* (*ethnicity*), which are non-modifiable. In addition, *heredity, genetic predisposition and inherited familial risks* are all very important, as well as *other diseases and conditions* (among others, *hypertension, high cholesterol levels, obesity* and *diabetes mellitus*);
- *behavioural risk factors* are associated with a person's daily choices, emotions and actions. Mainly they are inappropriate habits: *level of acute and/or chronic distress, dietary factors* (*eating customs, fat intake, alcohol consumption and excess*), *tobacco smoking, intake of multiple medications, level of physical activity* (often a *lack of physical exercises, sedentary behaviour*);
- *environmental risk factors* with regards to the interplay of individuals with their environment: *geo-graphic location, home hazards, hazardous features in the public environment, industrial toxins and poisons; the chaotic technological development of civilization;*
- *socio-economic risk factors* connected with a person's social conditions and the economic status of the individual which has a direct impact on access to healthcare: *occupation, social status, other social determinants of health* (*poor housing, low education, low degree of social interaction, low income, limited access to social healthcare services*).

Some examples of risk factors connected to a specific disease in the second part of life and in the ageing period:

- cardiovascular diseases: heredity (genetic factors); other diseases (obesity, hypertension, diabetes); stressful lifestyle; smoking; wrong and harmful diet habits (not drinking enough water, too much salt in the diet, increased fate and/or sugar intake, high LDL-cholesterol); lack of physical activity and exercises; drug use, abuse and combinations;
- stroke: advanced age; hypertension; previous stroke or transient ischemic attacks; diabetes; high LDL-cholesterol levels; smoking;
- Alzheimer's disease: advancing age; family history and heredity (risk and deterministic genes);
- complex interactions among genes and other risk factors, resulting from defective lifestyle and the deficient management of health conditions (i.e. head trauma, heart-brain connection and pathology, cardiovascular risk factors, interference with vascular dementia, low level of education);
- breast cancer: age, gender and racial factors; heredity (BRCA 1 and BRCA 2 autosomal dominant genes), prior cancers, hormones and obesity; dietary factors; environmental chemical and physical agents; socioeconomic factors.

Controlling health risk factor, in relation to type, number and intensity, is paramount to the development of a global health strategy. Risk factors:

- are strong distressors;
- disturb the good functioning of human socioeconomic organization;
- increase the cost of healthcare services; and
- are taken into account as anti-globalization factors.

The impact of risk factors on health is represented by preclinical (infra-, sub-clinical) phases of disease, which are the chronic-silent periods. The action of risk factors, diseases of lifestyle and silent pathologies (e.g. hypertension, hyperglycaemia etc.) cumulate their negative effects and thus they self-amplify into cascades of diseases.

In the pre-senescence and pre-disease period of the individual, knowledge of the preclinical phase of disorders obliges one to perform sub-clinical diagnosis and evaluation, and as a consequence determines personalized prevention.

The preclinical diagnosis of ageing and disease involves the investigation of oxidative stress -inflammatory disorders by establishing a pre-morbid individual profile: assays of biomarkers for the oxidative stress - inflammation status (Cutler et al., 2005c).

The increase of oxidative damages (evaluated in blood/serum, urine and breath) and a decrease of protective/defence antioxidant capacity (in serum), together with the augmentation of inflammation markers (in serum) will lead over the course of time to changes in the proper state of differentiation (Cutler, 2005b): cancer, senescence and senility.

4. Construction of human health-longevity

4.1 Longevity health sciences and mental health. Common characteristics

LHS and MH are in essence a form of health promotion associated with preventative medicine. For this reason (S. Riga et al., 2009a):

- the complementarity of LHS ⇔ MH is evident as a binomial: the construction of one of them meaning the development of the other one and vice versa;
- the superposition of LHS with MH is total during the last cycles of life: mature adult → old adult → 3rd age (65-85 years) → 4th age (over 85 years);
- The dependence of LHS ⇔ MH coupled to ecology (human, social and environmental) is manifested antagonistically:
  - (–) in a negative register, *stress* ↔ *ageing tetrad*: aetio-pathogenic and morbigenerating factors, ways and processes;
  - (+) in a positive register, *health* ↔ *longevity tetrad:* resources, strategies and therapies for longevity and mental health.

Essentially, *bio-medical gerontology* is the global and interdisciplinary study of ageing phenomena in phylogeny, ontogeny and medicine, while *clinical gerontology* and *geriatrics* are the medicine of ill old people (consequences of senescence and senility). In opposition with geriatrics, *anti-ageing medicine* and *positive ageing* are causal and preventative (from childhood and adolescence). Therefore, anti-ageing medicine is focused on health and longevity development, in conformity with genetic programming, the theoretical estimate of the maximum human lifespan being around 125 years (Weon and Je, 2009). *Longevity health sciences* and *SENS (Strategies for Engineered Negligible Senescence*) involve the utilization of advanced studies and translational medicine in public policies, in health and longevity (causes, resources, means, evaluations, programs and strategies) (D. Riga, 2003; de Grey, 2004; D. Riga and S. Riga, 2007).

4.2 From health to health-longevity

Man is a bio-psycho-social being, in close interrelation with his environment. Therefore, the bio-psycho-socio-ecological dimension of contemporary humans is fundamental for health-longevity (S. Riga et al., 2010c).

On the other hand, the 1946 WHO definition of health (*a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity)* confirms the bio-psycho-socio-ecological determinant of contemporary man. The definition of healthy ageing (Haber, 2003) comprises the following three components:

- *health promotion,* which includes strategies for reducing lifestyle risk factors as well as concepts for increasing healthy lifestyle habits;
- *health protection,* which contains strategies for modifying social and environmental structural health risks;
- *disease prevention,* which includes strategies to maintain and to improve health through medical care systems.

At present, the percentage of determining factors in ensuring health is as follows:

- lifestyle - 51%;
- biologic factor - 20%;
- environment - 19%;
- health care system - 10%.

Their control at national, regional and global levels involves coherent and efficient measures and strategies.

4.3 Palaestric civilization

The concept of palaestric civilization is an integrative global health conception (D. Riga and S. Riga, 2010a). At present, it comprises the beliefs, customs and culture of the ancients, the Renaissance ideals of physical beauty attained through exercise, the 19th - 20th Century efforts to institutionalize, generalize and popularize physical education and sports, and contemporary strategies of complementary health nutrition-physical activity.

The palaestric principles, characteristics which are clearly defined and highly positive, are:

- applicable throughout ontogenesis: child, adolescent, adult, old person;
- universal, efficient, long-term, easily put into practice, pleasant (entertaining) and low-cost;
- sanogenetic-prophylactic, therapeutic and recuperative (Bogdan and Bogdan, 2009);
- entropic, reorganizing, physical and cerebral activator, motivating, volitive, re-balancing (D. Riga and S. Riga, 2007).

The palaestric remedies work quite efficiently owing to the strong, long-term, multiple, positive effects that daily physical activity displays. Thus, they are important factors in:

- anti-stress, by lowering distress and raising eustress;
- anti-impairment, against the negative effects of daily life: lack of utilization, socio-sensory deprivation and physical inactivity, which is a complex deprivation, namely socio-sensory-effector deprivation (tactile, exercise and physical activity deprivation) and by overwork;

- anti-senescence, since they are somatic and psychic ageing decelerators;
- anti-polypathology, resulting from sedentariness and dysmetabolic syndrome: muscular atrophy, joint stiffness, osteoporosis, obesity, high blood pressure, diabetes, cardiovascular diseases, chronic fatigue syndrome.

There is a positive correlation between nourishment and exercise. Both rational nutrition and regular physical activity contribute to maintaining and improving good health (Simopoulos, 2005). Moreover, the palaestric solution also takes into account the bio-psycho-socio-ecological human dimension (S. Riga et al., 2010c). Physical education is a contributing factor in biologically and socially harmonizing a human being, as well as in integrating humans in their natural surroundings. In palaestric education, healthy nutrition is the 1st strategy for health-longevity. An unhealthy diet represents a major risk factor in non-communicable/chronic diseases, in the causation of global morbidity and for mortality. A lifestyle including physical activity is the 2nd principle and remedy. Physical inactivity represents a pathological habit, which increases the prevalence of 25 chronic diseases and produces more than 2 million deaths worldwide.

At present, there is strong global concern in relation to educating individuals in view of leading a healthier lifestyle, irrespective of age. In this sense, the palaestric paradigm, scientifically backed up by a large number of studies and researches, is prefigured as a valid solution. *The Declaration of Olympia,* May 28-29, 1996, drawn out and published one hundred years after 1896, when the modern and contemporary Olympic games were resumed in Athens, and the *WHO Documents and Recommendations* and the *European Union Legislation* (*White paper on a Strategy for Europe on Nutrition, Overweight and Obesity related Health Issue,* 2007; *White paper on Sport,* 2007) officially advocate the necessity of physical culture and education for each individual, as well as for the entire human society.

4.4 Declaration of Olympia on nutrition and fitness

4.4.1 Ancient Olympia, Greece, May 28-29, 1996 (Simopoulos, 2005)

1. Nutrition and physical activity interact in harmony and are the two most important positive factors that contribute to metabolic fitness and health interacting with the genetic endowment of the individual. Genes define opportunities for health and susceptibility to disease, while environmental factors determine which susceptible individuals will develop illness. Therefore, individual variation may need to be considered to achieve optimal health and to correct disorders associated with micronutrient deficiency, dietary imbalance and a sedentary lifestyle.
2. Every child and adult needs sufficient food and physical activity to express their genetic potential for growth, development, and health. Insufficient consumption of energy, protein, essential fatty acids, vitamins (particularly vitamins A, C, D, E and the B complex) and minerals (particularly calcium, iron, iodine, potassium and zinc), and inadequate opportunities for physical activity impair the attainment of overall health and musculoskeletal function.
3. Balancing physical activity and good nutrition for fitness is best illustrated by the concept of energy intake and output. For sedentary populations, physical activity must be increased; for populations engaging in intense occupational and/or recreational physical activities, food consumption may need to be increased to meet their energy needs.

4. Nutrient intakes should match more closely human evolutionary heritage. The choice of foods should lead to a diverse diet high in fruits and vegetables and rich in essential nutrients, particularly protective antioxidants and essential fatty acids.
5. The current level of physical activity should match more closely our genetic endowment. [The] reestablishment of regular physical activity into everyday life on a daily basis is essential for physical, mental and spiritual well-being. For all ages and both genders the physical activity should be appropriately vigorous and of sufficient duration, frequency, and intensity, using large muscle groups rhythmically and repetitively. Special attention to adequate nutrition should be given to competitive athletes.
6. The attainment of metabolic fitness through energy balance, good nutrition and physical activity reduces the risk of and forms the treatment framework for many modern lifestyle diseases such as diabetes mellitus, hypertension, osteoporosis, some cancers, obesity, and cardiovascular disorders. Metabolic fitness maintains and improves musculoskeletal function, mobility, and the activities of daily living into old age.
7. Education regarding healthy nutrition and physical activity must begin early and continue throughout life. Nutrition and physical activity must be interwoven into the curriculum of school age children and of educators, nutritionists and other health professionals. Positive role models must be developed and prompted by society and the media.
8. Major personal behavioural changes supported by the family, the community, and societal resources are necessary to reject unhealthy lifestyles and to embrace an active lifestyle and good nutrition.
9. National governments and the private sector must coordinate their efforts to encourage good nutrition and physical activity throughout the life cycle and thus increase the pool of physically fit individuals who emulate the Olympic ideal.
10. The ancient Greeks (Hellenes) attained a high level of civilization based on good nutrition, regular physical activity, and intellectual development. They strove for excellence in mind and body. Modern men, women, and children can emulate this Olympic ideal and become swifter, stronger and fitter through regular physical activity and good nutrition".

4.5 New conception - strategy - therapeutics in pro-longevity medicine

Anti-stress, anti-impairment, anti-ageing and anti-pathology therapy is a new specific, simultaneous and synergistic strategy and conception in preventative, curative and recovery medicine (Class of the Antagonic-Stress® drugs), (D. Riga and S. Riga, 1995-2005).

The therapy acts aetio-pathogenically in antagonizing and attenuating the *stress ↔ ageing tetrad (mental-biologic-oxidative-inflammatory distress ↔ impairment-wear and tear ↔ normal and accelerated ageing-inflammaging ↔ poly-pathologies as stress- and age-associated diseases),* at metabolic, subcellular, cellular, tissual, organic and systemic levels. This way, the entropic cascade of *stress ↔ ageing* is replaced with the *health ↔ longevity,* anti-entropic, protective-therapeutic *tetrad*: anti-stress/eustress/adaptation ↔ anti-impairment/vitality/resistance ↔ anti-ageing/active, healthy longevity ↔ anti-illnesses/anti-diseases. In addition, this first-hand restorative therapy recovers the anti-oxidative capacity/reserve/defence, a feature of the human body which has a direct relation with health-longevity.

The drug-therapy was elaborated by association of the following active principles:

- against oxidative and catabolic stress: methionine with aminoethanol phenoxyacetates and/or aminoethyl phenoxyacetamides;
- against anabolic stress: hydrooxopyrimidine carboxylates and/or oxopyrrolidine acetamides with potassium, zinc and lithium;
- vasodilative and normolipidemic: nicotinic alcohol and/or acid, or its derivatives, with magnesium and iodine;
- energo-active and anti-toxic: aspartate, fructose, vitamin B1, vitamin B6, monoacid phosphate and sulphate.

The process for manufacturing the drug stipulates:

- pharmaceutical preparation in two complementary types of capsules or coated tablets, gastrosoluble and enterosoluble, the last being enteric coated;
- prolonged-release of vasodilator from the enterosoluble unit.

For competitive and long-term health-longevity, this original therapy must be associated and integrated with:

- healthy diet, nutraceuticals, and regenerative bioactive factors;
- caloric restriction with adequate nutrition;
- cerebral activation therapy, other antioxidants, nootropics, neurovascular and neurometabolic activators;
- cognitive stimulation, continuous learning-education, brain training and fitness;
- regular exercise, daily physical activity, and resistance exercises;
- hormesis, including adaptation to stimulation, and low-level stress (Rattan and Demirovic, 2009).

5. Health-longevity strategy

5.1 Quality of life for all. The WHO public health policy

"Targets for Health for All - 2000" is a global strategy envisioned by the WHO and represents a programmatic document (WHO, Regional Office for Europe, 1986): "Primary health care is the most important single element in the reorientation of the health care system and will require very strong support" (p. 11). For this objective, "Lifestyles conductive to health" (Ch. 4) and a "Healthy environment" (Ch. 5) become fundamental.

The six important subjects and the four dimensions of health promotion were very well emphasized:

"Six major themes run throughout the whole book.

- Health for all implies *equity.* This means that the present inequalities in health between countries and within countries should be reduced as far as possible.
- The aim is to give a positive sense of health so that they can make full use of their physical, mental and emotional capacities. The main emphasis should therefore be on *health promotion* and the prevention of disease.
- Health for all will be achieved by people themselves. A well-informed, well-motivated and actively *participating community* is a key element for the attainment of the common goal.

- Health for all requires the coordinated action of all sectors concerned. The health authorities can deal only with a part of the problems to be solved, and *multisectoral cooperation* is the only way of effectively ensuring the prerequisites for health, promoting healthy policies and reducing risks in the physical, economic and social environment.
- The focus of the health care system should be on *primary health care* - meeting the basic health needs of each community through services provided as close as possible to where people live and work, readily accessible and acceptable to all, and based on full community participation.
- Health problems transcend national frontiers. Pollution and trade in health-damaging products are obvious examples of problems whose solution requires *international cooperation*" (pp. 5-6).

"Thus, health for all in Europe has four dimensions as regards health outcomes, involving action in order to:

- *ensure equity in health,* by reducing the present gap in health status between countries and groups within countries;
- *add life to years,* by ensuring the full development and use of people's integral or residual physical and mental capacity to derive the full benefit from it and to cope with life in a healthy way;
- *add health to life,* by reducing disease and disability;
- *add years to life,* by reducing premature deaths, and thereby increasing life expectancy" (p. 23).

The WHO (a specialized agency of the United Nations, primarily responsible for international public health) published, in 1987, an essential tool: "Measurement in health promotion and protection" (Abelin et al., 1987). This WHO manual represents a new health movement for a global strategy, promoting positive health, in the socio-ecological paradigm of health. Therefore, "the main goal of health promotion is to maintain or improve health potential" (p. 19).

Also, on October 12, 1990, the WHO teleconference cautions against "diseases of lifestyle", which are the cause of 70-80% of premature deaths in industrialized countries. Thus, health promotion signifies the prevention of stress-related diseases (Cooper, 1996).

Therefore, the quality of life for all represents the promotion of positive health, a new socio-ecological paradigm of health and preventative medicine (S. Riga and D. Riga, 2009b).

5.2 Health ↔ longevity tetrad

Mental (psychic, behavioural) and somatic (body, metabolic) health with the construction of the health-longevity couple represent the medicine of the future. The *health ↔ longevity tetrad (anti-stress ↔ anti-impairment ↔ anti-ageing ↔ anti-diseases)* is in total opposition with the stress ↔ ageing cascade.

LHS and MH have common principles and strategies. Both:

- will reform the previous paradigm of contemporary medicine (**Figure 3**), the modern pyramid of (mental) medical services (Funk et al., 2007), from treatments and illness recovery;

- to the medicine of the healthy individual (**Figure 4**), - New pyramid of (mental) health services (S. Riga et al., 2009a; S. Riga et al., 2011a).

Fig. 3. Modern pyramid of (mental) medical services.
Optimal mix recommended by WHO (2007)

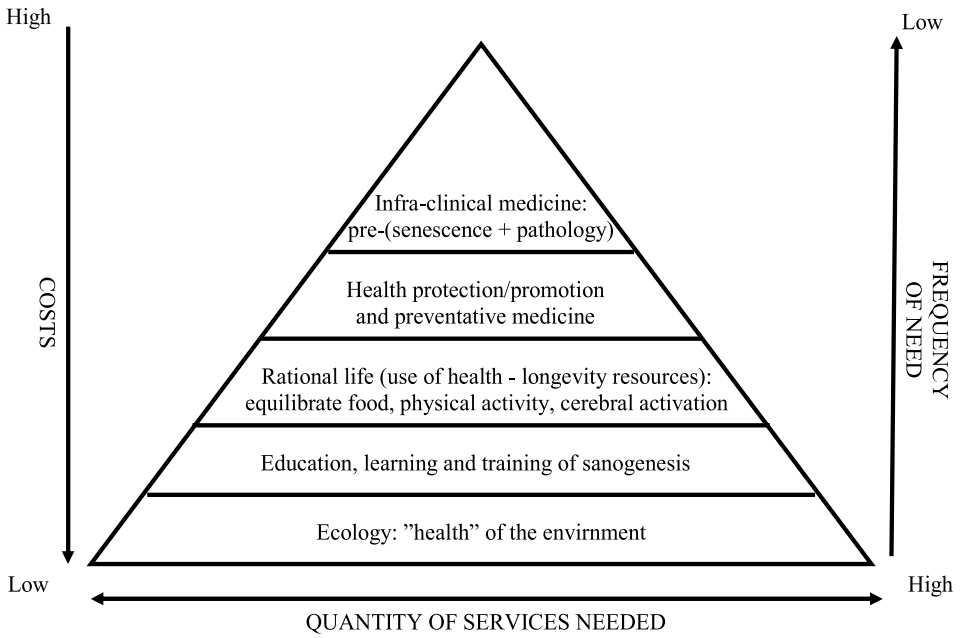


Fig. 4. New pyramid of (mental) health services.
Advanced paradigm in (mental) health - longevity services (2009)

The societal cost/benefit ratio is decisively in favour of health-longevity promotion, in comparison with current medical care systems, represented by polyclinics, hospitals and sanatoriums. The cost/benefit ratio will always rank prevention and prophylaxis as higher place than therapeutics and recovery whenever savings and economic factors are involved.

5.3 New health-longevity strategy. Structure of health as a pyramid

This original paradigm is structured in *a new pyramid* of health-longevity services (S. Riga and D. Riga, 2009a; S. Riga et al., 2011a), with five levels:

1. Ecology: "the health" of the environment, permanent human healthy conceptions and actions on the surroundings, normal human-environment interactions;
2. The culture of sanogenesis, which involves education, learning, construction, development, training, maintenance, continuity and permanence;
3. Rational life and use of health-longevity resources: balanced diet and often dietary restriction, regular physical activity, cerebral metabolic activation, cognitive and social stimulation, hormesis;
4. Health protection (promotion) and preventative medicine;
5. Sub-clinical (infra-clinical) medicine, with developmental origins of health and diseases, risk factors for health, biologic and psychic impairment, pre-senescence, pre-illness and silent pathologies.

An optimal mix of ecological, bio-medical and care systems and services in the promotion of health-longevity integrates the costs (left side), the frequency of needs (right side) and the quantity of services needed (presented on a horizontal line). The most favourable and viable combination is structured as a new pyramid of health-longevity services (**Figure 4**), (S. Riga et al., 2010d; S. Riga et al., 2011a).

From the base to the top, the hierarchy of services needed comprises five levels:

1. Ecology: the "health" of the environment (natural, artificial, societal, regional and, finally, global - the earth), (WHO, Regional Office for Europe, 1986; Abelin et al., 1987);
2. The continuous education, learning and training of sanogenesis (Abelin et al., 1987; S. Riga et al., 2009b): 1st stage (cognitive education → construction → development) and 2nd stage (maintenance → training/coaching → improvement → continuity / permanence);
3. The rational utilization of personal life and health-longevity resources (Klatz and Goldman, 2003; Le Bourg, 2003; Simopoulous, 2005; D. Riga et al., 2006b): diet, physical activity, cerebral activation (psychic, nutraceutical, metabolic, psychological and social);
4. Health protection → promotion → development and preventative medicine (primary prophylaxis), (WHO, Regional Office for Europe, 1986; Abelin et al., 1987; Knapp et al., 2007; S. Riga and D. Riga, 2008);
5. Infra-clinical medicine in pre-senescence and pre-pathology (Cutler, 1996; Cutler et al., 2005a; D. Riga and S. Riga, 2007): diagnosis - evaluation - intervention for risk factors, inductors of pre-senescence, pre-illness and silent pathology and, finally, for diseases (markers of oxidative stress and inflammation, cancer antigens etc.).

5.4 Health-longevity - A global progress

The First Law *(Law of use and disuse),* in its extended form, enunciated by Jean-Baptiste Lamarck (1744-1828), the French naturalist, is very important for the health-longevity strategy: *In every animal which has not passed the limit of its development, a more frequent and continuous use of any organ gradually strengthens, develops and enlarges that organ, and gives it a power proportional to the length of time it has been so used; while the permanent disuse of any organ imperceptibly weakens and deteriorates it, and progressively diminishes its functional capacity, until it finally disappears* (Lamark, 1809, trans. 1914).

As an actual concept, it becomes "use it or lose it" (engl.)/"utilisez-la ou perdez-la" (fr.), both for neurons (Swaab, 1991) as well as for mental activity (Roth, 1975; Giurgea, 1993), namely therapy for cerebral activation, utilized in sanogenesis, prophylaxis of neuro-degenerative diseases and against pathological ageing.

At an individual (personalized) level the continuous education of health is defining.

At a national (societal) level, for an increased efficacy of health-longevity strategies, two directions must be covered:

- the improvement of programmes for the assessment of risks of diseases and of the precocious discovery of illnesses, followed by:
- the elaboration and implementation of programmes for health-longevity improvement and maintenance.

Now, is the time to create global standards in the training of health promotion. For this reason, the International Institute for Health promotion was organized in 1996 at the American University in Washington, DC (Kirsten, 2010), as an interdisciplinary network of specialists from various fields, and also of academic, governmental and non- governmental organizations.

In our new conception, the aim of health-longevity is health promotion together with illness prevention and the improvement of the quality of life. Moreover, the advantages of the proposed public health strategies and policies (pyramid of health) are low societal costs compared to the enduring treatments for chronic diseases. Therefore, a new millennium strategy for a healthy person's medicine must entail qualified interventions:

- in the early life of the origins of human health and disease (Newnham and Ross, 2009);
- in stress-ageing aetio-pathogenic entropic cascade (distress-impairment-ageing-illness), (Fahy et al., 2010; D. Riga et al., 2006a; D. Riga et al., 2006b; S. Riga et al., 2010d);
- in diseases of lifestyle, risk factors, silent pathologies (persistent mental - biologic - oxidative - inflammatory stress).

Consequently, future medicine will be and must be the medicine of health, mainly the planning of personalized and public health, together with the strategies of longevity, somatic and mental health.

The ageing of the population (implicitly chronic diseases) and also mental/behavioural disorders are in rapid expansion. Due to the high public costs, these phenomena will force society towards a new health policy: health protection/promotion and preventative/prophylactic medicine. Consequently, in the global world, the future medicine

will be the medicine of health: the planning of personalized/public health and strategies of longevity/mental health.

In 2002, non-communicable diseases accounted for 60% of total mortality worldwide and 46% of the global burden of disease (WHO, 2003). This disease burden is expected to increase from 46% in 2002 to 60% in 2020. The major causes of this are represented by five factors (high blood pressure, high cholesterol, low intake of vegetables and fruits, high body mass index and physical inactivity) from the top 10 global disease burden factors enumerated by the WHO. These current risk levels (a worldwide risk diagram) predict major increases in chronic diseases, as a poly-pathology of ageing.

On May 2004, at the 56th World Health Assembly, the WHO substantiated an important global public health initiative (Waxman, 2005), the main targets of which were diet, physical activity and health.

6. Conclusions

The progress in science, medicine, technology and communication imposes global policies - strategies - standards in health promotion from the WHO regarding education, training, expertise, culture and research.

Contemporary civilization should therefore substantiate key competences:

- durable health development;
- a knowledge-based society;
- social, communication and civic abilities;
- learning to learn competencies.

Health-longevity medicine is a new concept for public health, health promotion and protection, in accordance with world demographic tendencies. This strategy for future health at a global level reunites preventative (prophylaxis and hygiene) medicine, LHS, MH and the human bio-psycho-socio-ecological dimension.

7. References

[1] Abelin T., Brzezinski Z. J., Carstairs V. D. L., Eds. 1987. *Measurement in Health Promotion and Protection.* WHO Regional Publications, European Series No. 22. World Health Organization, Regional Office for Europe. Copenhagen, DK.

[2] Bogdan V., Bogdan A. 2009. The sanogenetic role of physical activity. Why should we wait until it is too late? *Palestrica of the 3rd Millennium - Civilization and Sport.* 10: 48-53.

[3] Commission of the European Communities. 2007. *White Paper on a Strategy for Europe on Nutrition, Overweight and Obesity related Health Issue*. http://ec.europa.eu/health/ ph_determinants/life_style/nutrition/documents/nutrition_wp_ ro.pdf

[4] Commission of the European Communities. 2007. *White Paper on Sport.* http://ec.europa.eu/ sport/whitepaper/wp_on_sport_ro.pdf

[5] Cooper C. L., Ed. 1996. *Handbook of Stress, Medicine and Health.* CRC Press. Boca Raton, FL.

[6] Cutler R. G. 1996. The molecular and evolutionary aspects of human aging and longevity. In: *Advances in Anti-Aging Medicine.* R. Klatz, Ed.: 71-99. Mary Ann Liebert. New York, NY.

[7] Cutler R. G., Harman S. M., Heward C., Gibbons M., Eds. 2005a. *Longevity Health Sciences. The Phoenix Conference.* Ann. N. Y. Acad. Sci, Vol. 1055. New York Academy of Sciences. New York, NY.

[8] Cutler R. G. 2005b. Oxidative stress profiling. Part I. Its potential importance in the optimization of human health. *Ann. N. Y. Acad. Sci.* 1055: 93-135.

[9] Cutler R. G., Plummer J., Chowdhury K., Heward C. 2005c. Oxidative stress profiling. Part II. Theory, technology and practice. *Ann. N. Y. Acad. Sci.* 1055: 136-158.

[10] de Grey A. D. N. J., Ed. 2004. *Strategies for Engineered Negligible Senescence. Why Genuine Control of Aging May Be Foreseeable.* Ann. N. Y. Acad. Sci, Vol. 1019. New York Academy of Sciences. New York, NY.

[11] Fahy G. M., West M. D, Coles L. S., Harris S. B., Eds. 2010. *The Future of Aging. Pathways to Human Life Extension.* Springer. Dordrecht, NL.

[12] Funk M., Drew N., Saraceno B. 2007. Global perspective on mental health policy and service development issues: the WHO angle. In: *Mental Health Policy and Practice across Europe. The future direction of mental health care.* M. Knapp, D. McDaid, E. Mossialos & G. Thornicroft, Eds.: 426-440. Open University Press. New York, NY.

[13] Giurgea C. E. 1993. *Le vieillissement cérébral normal et réussi. Le défi du XXI^e siècle.* Mardaga. Liège, BE.

[14] Haber D. 2003. *Health Promotion and Aging. Practical Applications for Health Professionals,* 3rd ed. Springer. New York, NY.

[15] Hanson A. E. 2006. *Hippocrates: the "Greek Miracle" in Medicine.* L. T. Pearcy, The Episcopal Academy. Merion, PA.

[16] Harman D. 1984. Free radical theory of aging: the "free radical" diseases. *Age.* 7: 111-131.

[17] Klatz R., Goldman R. 2003. *The New Anti-Aging Revolution.* Basic Health Publ. North Bergen, NJ.

[18] Knapp M., McDaid D., Mossialos E., Thornicroft G., Eds. 2007. *Mental Health Policy and Practice Across Europe.* Open University Press. McGraw-Hill Education. Maidenhead, UK.

[19] Kristen W. 2010. Creating global standards in health promotion training - the International Institute for Health Promotion. *Palestrica of the 3rd Millennium - Civilization and Sport.* 11: 291-292.

[20] Lamarck J.-B. 1914. *Philosophie zoologique, ou exposition des considérations relatives à l'histoire naturelle des animaux, 1809.* Trans. by H. Elliot. Macmillan. London, UK.

[21] Le Bourg E. 2003. Antioxidants as Modulators. In: *Modulating Aging and Longevity.* S. I. S. Rattan, Ed.: 183-203. Kluwer. Dordrecht, NL.

[22] Lin H. B. 2000. *Chinese Health Care Secrets. A Natural Lifestyle Approach.* Llewellyn Publ., St. Paul, MN.

[23] Miwa S., Beckman K. B., Muller F. L., Eds. 2008. *Oxidative Stress in Aging. From Model Systems to Human Diseases.* Humana Press-Springer Science. Totowa, NJ.

[24] Muller D. P. R., Goss-Sampson M. A., MacEvilly C. J. 1992. Antioxidant deficiency and neurological disease in humans and experimental animals. In: *Free Radicals in the Brain. Aging, Neurological and Mental Disorders.* L. Packer, L. Prilipko & Y. Christen, Eds.: 62-73. Springer. Berlin, DE.

[25] Newnham J. P., Ross M. G., Eds. 2009. *Early Life Origins of Human Health and Disease.* Karger. Basel, CH.

[26] Rattan S. I. S., Demirovic D. 2009. Hormesis and aging. In: *Hormesis: a revolution in biology toxicology and medicine.* M. P. Mattson & E. Calabrese, Eds.: 153-175. Springer. New York, NY.

[27] Riga D., Riga S. 1995-2005. *Anti-stress, anti-impairment and anti-aging drugs and process for manufacturing thereof (Class of the Antagonic-Stress® drugs/therapy - Dr. Dan Riga & Dr. Sorin Riga),* 64 pp., a new conception - strategy - therapeutics with 27 worldwide patents in 3 international organizations, 25 states and 5 continents: WIPO, PCT: PCT/WO 95/33486;

[28] EPO: EUR. Pat. 1999 (17 countries - AT, BE, CH, DE, DK, ES, FR, GB, IE, IT, LI, LU, MC, NL, PT, RO, SE);

[29] AU Pat. 1998, KR Pat. 1999, RU Pat. 2000, US Pat. 2001, CN Pat. 2001, CA Pat. 2002, JP Pat. 2003, BR Pat. 2005.

[30] Riga D. 2003. SENS acquires SENSe: present and future anti-aging strategies. *J. Anti-Aging Med. (Rejuvenation Res.)* 6: 231-236.

[31] Riga D., Riga S., Schneider Fr. 2004a. Regenerative medicine: Antagonic-Stress® therapy in distress and aging. I. Preclinical synthesis - 2003. *Ann. N.Y. Acad. Sci.* 1019: 396-400.

[32] Riga S., Riga D., Schneider Fr. 2004b. Prolongevity medicine: Antagonic-Stress® drug in distress, geriatrics and related diseases. II. Clinical review - 2003. *Ann. N.Y. Acad. Sci.* 1019: 401-405.

[33] Riga D., Riga S., Hălălău F., Schneider Fr. 2006a. Lipofuscin and ceroid pigments - markers of normal and pathological brain aging. In *Anti-Aging Therapeutics,* Vol. 8, R. Klatz, R. Goldman, Eds.: 213-221. American Academy of Anti-Aging Medicine. Chicago, IL.

[34] Riga D., Riga S., Hălălău F., Schneider Fr. 2006b. Neurono-glial mechanisms in brain protection, aging deceleration and neuro-psycho-longevity. In *Anti-Aging Therapeutics,* Vol. 8, R. Klatz, R. Goldman, Eds.: 223-236. American Academy of Anti-Aging Medicine. Chicago, IL.

[35] Riga D., Riga S. 2007. *Anti-Aging Medicine and Longevity Sciences* (Romanian lang.). Cartea Universitara Publ. Bucharest, RO.

[36] Riga S., Riga D. 2008. *Stressology, Adaptology and Mental Health* (Romanian lang.). Cartea Universitara Publ. Bucharest, RO.

[37] Riga S., Riga D., Danailă L., Mihăilescu A., Motoc D., Moş L., Schneider, Fr. 2009a. New politics for global health and longevity: complementarity of anti-aging medicine with mental health. *19th IAGG (Int. Assoc. Gerontol. Geriatrics) World Congress.* July 5-9, 2009. Paris, FR. *Abstract* PB7 495. *J. Nutr. Health Aging.* 13(S1): S475.

[38] Riga S., Riga D., Danăilă L., Mihăilescu A., Motoc D., Moş L., Schneider, Fr. 2009b. Longevity science and mental health - unification of their concepts and strategies, essential key for the future medicine. *13th IABG (Int. Assoc. Biomed. Gerontol.) Congress.* Québec, CA, May 18-20, 2009.

[39] Riga D., Riga S., Moş L., Motoc D., Schneider Fr. 2009c. Pro-longevity life styles. Importance of physical activity and sport. *Palestrica of the 3rd Millennium - Civilization and Sport.* 10: 138-144.

[40] Riga D., Riga S. 2010a. Palestra's paradigm. *Palestrica of the 3rd Millennium - Civilization and Sport.* 11: 7-9.

[41] Riga D., Riga S., Ardelean A., Schneider Fr. 2010b. Health, longevity and ecology - an integrated paradigm. *Fiziologia-Physiology,* 20(1): 13-16.

[42] Riga S., Riga D., Ardelean A., Schneider Fr. 2010c. The contemporary man in his bio-psycho-socio-ecological dimension. *Fiziologia-Physiology,* 20(2): 8-10.

[43] Riga S., Riga D., Mihăilescu A., Motoc D., Moş L, Schneider Fr. 2010d. Longevity health sciences and mental health as future medicine. *Ann. N.Y. Acad. Sci.* 1197: 184-187.

[44] Riga S., Riga D., Ghinescu M., Mihăilescu A., Motoc D., Geacăr S. 2011a. Health-Longevity Pyramid in the Anti-Aging Global Progress. *61st Annual Scientific Meeting of the British Society for Research on Ageing - BSRA & 14th Congress of the International Association of Biological Gerontology - IABG,* Brighton, UK, July 11-14, 2011.

[45] Riga D., Riga S. 2011b. The Science of Ageing - Global Progress, *Rejuvenation Res.* 14: 573-577.

[46] Roth M. 1975. The diagnosis of dementia. *Br. J. Psychiatry.* 125(9): 87-99.

[47] Selye H. 1976. Stress in Health and Disease. Butterworths. Boston, MA.

[48] Simopoulous A. P., Ed. 2005. *Nutrition and Fitness.* Vol. 1 - *Obesity, the Metabolic Syndrome, Cardiovascular Disease, and Cancer.* Vol. 2 - *Mental Health, Aging, and the Implementation of a Healthy Diet and Physical Activity Lifestyle.* Karger. Basel, CH.

[49] Slater T. F., Block G., Eds. 1991. Antioxidant vitamins and β-carotene in disease prevention. *Am. J. Clin. Nutr.* 53(S1): 189S-396S.

[50] Swaab, D. F. 1991. Brain aging and Alzheimer's disease, "wear and tear" versus "use it or lose it". *Neurobiol. Aging.* 12: 317-324.

[51] Unschuld P. U. 2003. *Huang Di nei jing su wen: Nature, Knowledge, Imagery in an Ancient Chinese Medical Text.* University of California Press, Berkeley, CA.

[52] Waxman A. 2005. Why a global strategy on diet, physical activity and health? In: *Nutrition and Fitness: Mental Health, Aging, and the Implementation of a Healthy Diet and Physical Activity Lifestyle.* Vol. 2. A. P. Simopoulous, Ed.: 162-166. Karger. Basel, CH.

[53] Weon B. M., Je J. H. 2009. Theoretical estimation of maximum human lifespan. *Biogerontology.* 10: 65-71.

[54] WHO, Regional Office for Europe. 1986. *Targets for Health for All - 2000. Targets in Support of the European Regional Strategy for Health for All.* World Health Organization, Regional Office for Europe. Copenhagen, DK.

[55] WHO. 1992. *The ICD-10. Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines.* World Health Organization. Geneva, CH.

[56] WHO. 2003. *Shaping the Future. The World Health Report*. World Health Organization. Geneva, CH.

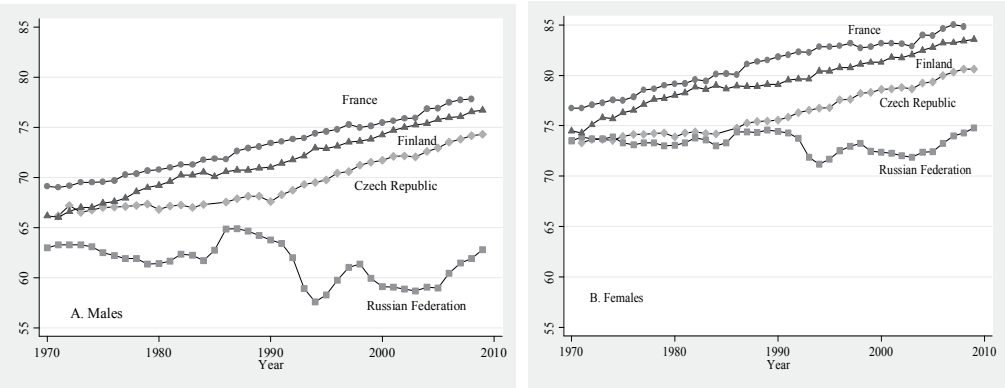
18

Alcoholism and the Russian Mortality Crisis

Irina Denisova[1] and Marina Kartseva[2]
[1]*New Economic School, Moscow*
[2]*Centre for Economic and Financial Research, Moscow*
Russia

1. Introduction

Life expectancy is the key aggregated indicator of a country's well-being along with gross domestic product and living standards. While Russia approaches the group of developed countries in terms of per capita GDP, it is strikingly different in terms of the living standards and the dynamics of life expectancy. Thus, life expectancy among males in Russia has not only not increased since the 1970s, but has dropped to barely above 60 years (Fig.1). The low living standards and lack of improvement in life expectancy dynamics in Russia are in contrast with the experience of the majority of developed countries and countries with transitional economies. Thus, male life expectancy at birth in Finland has increased from 66 years in 1970 to 76 years in 2007, in Norway from 71 to 77 years and in Sweden from 72 to 78 years during the same period. In the Czech Republic male life expectancy has increased from 66 to 68 years. Female life expectancy in these countries reveals comparable dynamics. Russia has still to go into an upward trend (both for men and for women) characteristic of all the developed and the majority of the developing countries.



Source: European Health for All Database, 2011.

Fig. 1. Life expectancy at birth, 1970-2007, male (left) and female (right). Top down: France, Finland, the Czech Republic, and the Russian Federation.

Numerous studies of the causes of high mortality among the Russian population all confirm the negative impact of excessive alcohol consumption (Leon et al, 1997; Shkolnikov et al,

1998; Brainerd and Cutler, 2005; Leon, 2007; Nemtsov, 2002). The majority of studies use aggregated death certificate data, which limits a more detailed study of the impact of alcohol consumption patterns on health and ultimately on the risk of death[1]. The data of the Russian Longitudinal Monitoring Survey (RLMS-HSE) make it possible to identify types of alcohol consumption and analyze the impact of the main types on health and the risk of death. Section 2 analyzes aggregate alcohol consumption in Russia and Europe. Section 3 is devoted to the structure of alcohol consumption in Russia. Section 4 reports the results of the assessment of the impact of alcohol consumption on health and mortality in Russia. Section 5 is devoted to the experience of European countries in implementing active anti-alcohol policies. Section 6 concludes.

2. Alcohol consumption in Russia and Europe

The total registered consumption of alcohol in Russia in 2008 reached 11.5 litres of pure alcohol per person above the age of 15 (Fig.2). The consumption of spirits increased by 233% between 1988 and 1998, the consumption of beer by 31%, while the consumption of wine dropped slightly by 6% (World Drink Trends, 1999, Global Status Report on Alcohol, 2004). The production of illicit (unregistered) products adds almost 5 litres, according to expert assessments[2]. That adds up to 16 litres of pure alcohol per citizen over 15 years of age (the Ministry of Healthcare and Social Development puts the figure at 18 litres). It has to be noted that unregistered alcohol consumption is not a peculiarly Russian phenomenon. In the majority of West European countries unregistered consumption is put by experts at between 5 and 20%, and sometimes at as much as a third of registered consumption.

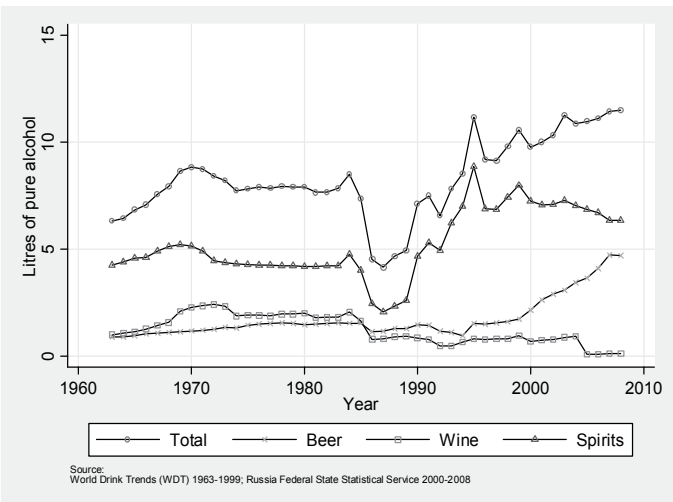


Fig. 2. Total (registered) consumption of alcohol in Russia, consumption of beer, spirits and wine. Litres of pure alcohol per capita of 15 + per year.

[1] A small group of studies are based on micro-data collected either with the express purpose of identifying the impact of harmful habits on the risk of death (surveys of the relatives of the dead in Izhevsk), or for other purposes (lipids test program).

[2] European Addiction Research (2001), Gilinskiy Y. (2000).

Aggregate consumption in Russia is higher than in the countries of Europe, although not much higher than in the Czech Republic, France and Germany. At the same time, the general trend of consumption in Europe and the US is a gradual decline in alcohol consumption (in litres of pure alcohol) since the 1980s, with the trend more manifest in Europe than in the US. Strong drinks are being replaced with lighter ones. Thus, for example, in the period between 1988 and 1998 consumption of spirits in Italy dropped by 50%, consumption of wine by 18% while consumption of beer increased by 15%. In Great Britain consumption of spirits and beer dropped by 28% and 17% respectively while wine consumption increased by a third (27%). During the same period in Europe as a whole consumption of spirits dropped by 23.2%, consumption of wine dropped by 3.6%, while consumption of beer increased by 3.6%[3]. Similar trends have been noted not only in developed countries. Thus in the majority of Latin American countries consumption of alcohol is going down and consumption of beer is going up.

For Russia the experience of North European countries where consumption patterns are historically similar to Russia is of the greatest interest. Figs. 3 and 4 show the dynamics of aggregate registered consumption of alcohol, beer, spirits and wine in Iceland, Finland, Norway and Sweden. As seen from the charts, all these countries witnessed dramatic changes in the structure of alcohol consumption in the 1980s and 1990s. The consumption of spirits dropped significantly: by 1.5 litres of pure alcohol per person in Norway, by 2 litres in Iceland and Finland and by almost 3 litres in Sweden. At the same time the total consumption of alcohol has not diminished and has actually grown a little because the consumption of spirits has been replaced with the consumption of beer and wine. As a result these countries moved from the group of countries with predominant consumption of strong spirits to countries with predominant consumption of beer.

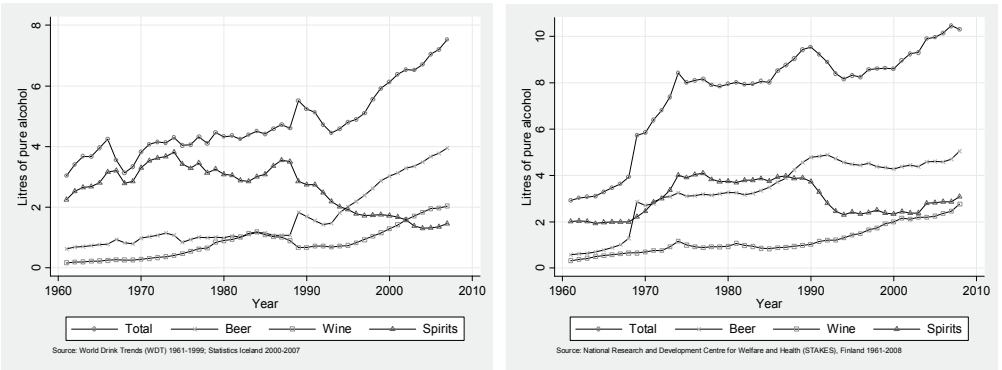


Fig. 3. Aggregate (registered) consumption of alcohol in Iceland (left) and Finland (right), consumption of beer, strong spirits and wine. Litres of pure alcohol per capita of 15 + per year.

[3] The calculations use data not from all the European countries, but from countries with larger populations (Belgium, Great Britain, France, Germany, Spain, Italy, Poland, the Czech Republic, Portugal and Switzerland).

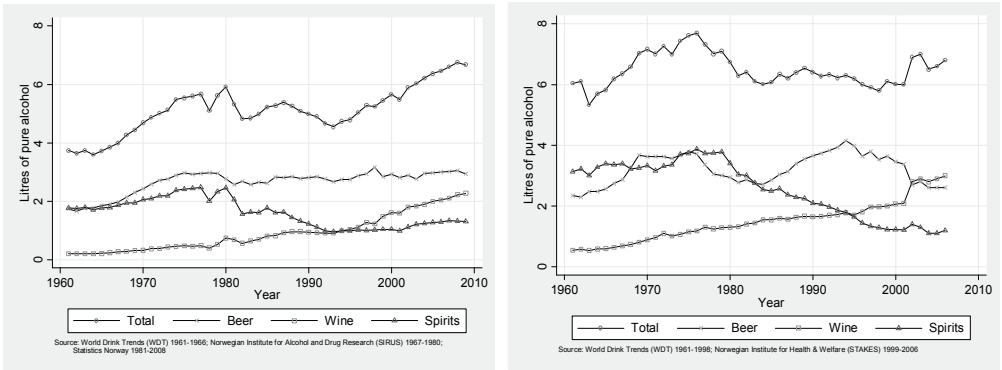


Fig. 4. Aggregate (registered) consumption of alcohol in Norway (left) and Sweden (right), consumption of beer, strong alcoholic beverages and wine. Litres of pure alcohol per capita of 15 + per year.

Changes in the structure of consumption, while not the only cause of increased life expectancy in the North European countries, have undoubtedly had a positive impact on bringing down the death rate and increasing life expectancy in these countries (see Fig.1). The change in the structure of alcohol consumption in Northern Europe has been the result of a massive, large-scale and sustained anti-alcohol policy in these countries. These measures will be discussed in more detail in Section 5.

It has to be noted that the switch from predominant consumption of spirits (hard liquor) to the consumption of beer or wine does not in itself guarantee lower risks of death. Another crucial factor is the frequency and volumes of alcohol consumption. Thus, France, which is traditionally a wine-drinking nation (Fig.5) has managed to reduce alcohol consumption almost by half between the 1960s and 2000 due to the reduction of wine consumption. The drop in consumption reduced the deaths from cardiovascular diseases (Fig.6). At the same time growing alcohol consumption, above all of spirits, in Russia has resulted in a growing death rate from these diseases.

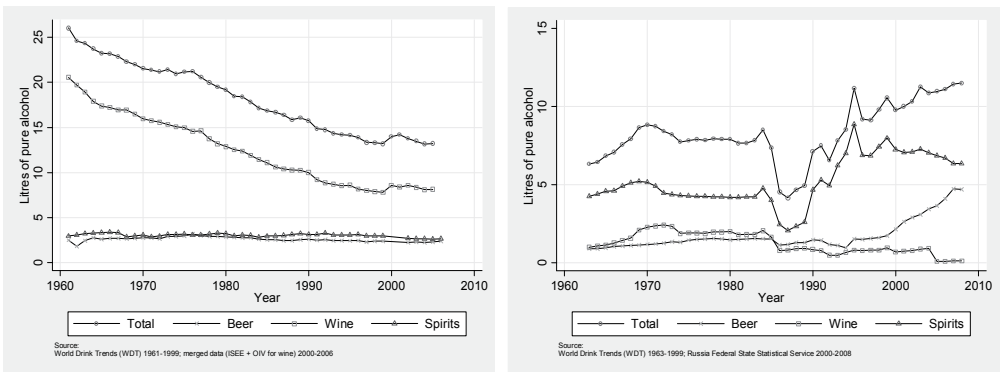
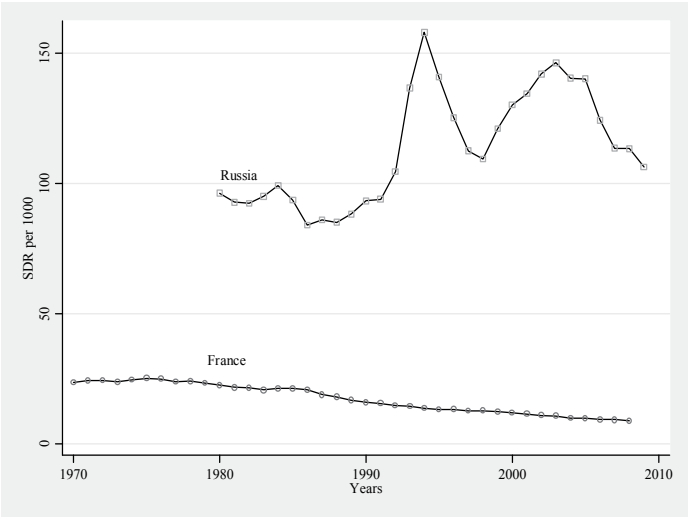


Fig. 5. Total (registered) alcohol consumption in France (left) and Russia (right), consumption of beer, strong alcoholic beverages and wine. Litres of pure alcohol per capita of 15 + per year.



Source: European health for all database, 2011

Fig. 6. Standardized coefficients of death from ischemic heart disease chronic diseases per 1000 persons in France and Russia.

On the whole aggregate alcohol consumption in Russia, although higher than in developed countries, is not so much higher as to explain the differences in mortality rate and life expectancy. It is true that the rate of alcohol-related deaths per litre consumed in Russia is substantially higher than similar indicators in Western Europe. The main reasons for that, as noted by scholars (e.g., Nemtsov, 2009) are the specific structure of consumption (a larger share of strong drinks), the northern type of alcohol consumption (large doses within a short time), the low standard of healthcare (especially the treatment of drug and alcohol addiction) as well as the traditional neglect of Russian people of their state of health.

3. Structure of alcohol consumption: frequency, volumes, beverages

As noted above, the type of consumption is a key characteristic of alcohol consumption (no less important than the amounts). The pattern of consumption is determined by the type of drinks in terms of strength and quality and the time and places when and where alcohol is consumed. Epidemiological studies in various countries show that the risk of cardiovascular diseases among those who drink a glass of wine a day is on average 32% less than among those who do not drink at all. A similar indicator for beer is 22% (Di Castelnuovo et al, 2002). Nemtsov (2009) notes that the impact of the pattern of consumption on the nation's health has been poorly studied by Russian narcologists. At the same time studies in other countries note that the "ideal structure" of alcohol consumption – the ratio that minimizes negative consequences – is consumption in which beer accounts for 50%, wine for 35% and spirits for 15% (Edwards et al., 1994). In 2002, according to official alcohol sales figures (that do not take into account illicit alcohol)

strong beverages accounted for 35% of the total consumption. The figure is obviously grossly understated because it does not take into account illicit production (both industrial and domestic).

The Russian Longitudinal Monitoring Survey (RLMS-HSE) makes it possible to analyze the structure of alcohol consumption by Russian households on the basis of respondents' answers[4]. According to the RLMS-HSE, about three quarters of the adult Russian population consume some kind of alcohol (Fig.7). The figure of drinkers is higher among males in all the age groups. The share of alcohol consumption is higher in the main age groups and a little lower in the 18-25 age group and among over 55s.

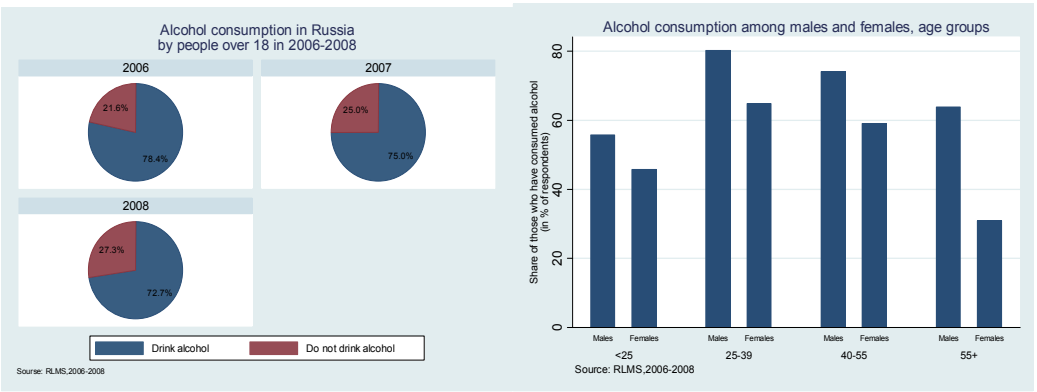


Fig. 7. Alcohol consumption in Russia[5]: total (left) and men and women by age groups (right). Percentage of total number of respondents.

It has to be noted that the share of alcohol drinkers is higher in groups with higher incomes and among those who have finished vocational training schools (PTU) or have higher education, although the difference in consumption depending on education is not great (Fig.8).

[4] RLMS-HSE is a nationally representative longitudinal survey of Russian households conducted since 1992 by the Demoscope Centre, the RAS Sociology Institute and the University of North Carolina at Chapel Hill (USA) Population Center. The National Research University Higher School of Economics (Moscow) joined the group in 2008. Cooperation with the top world centers for the study of the behavior of households in forming the sample, developing the questionnaire, recruiting and training interviewers earned this study a high degree of trust among Russian and foreign scholars and decision-makers. The RLMS-HSE data are nationally representative and are based on a survey of more than 4,000 households per year which amounts to more than 10,000 adults per year. The sample is from a two-stage random draw of dwellings from the population from the micro census of 1989. The dwellings are surveyed each year with some additional dwellings added in the later periods of the survey to meet the national representation criteria (http://www.cpc.unc.edu/projects/rlms-hse).

[5] Respondents were asked whether they had consumed any alcohol (including beer) in the last 30 days.

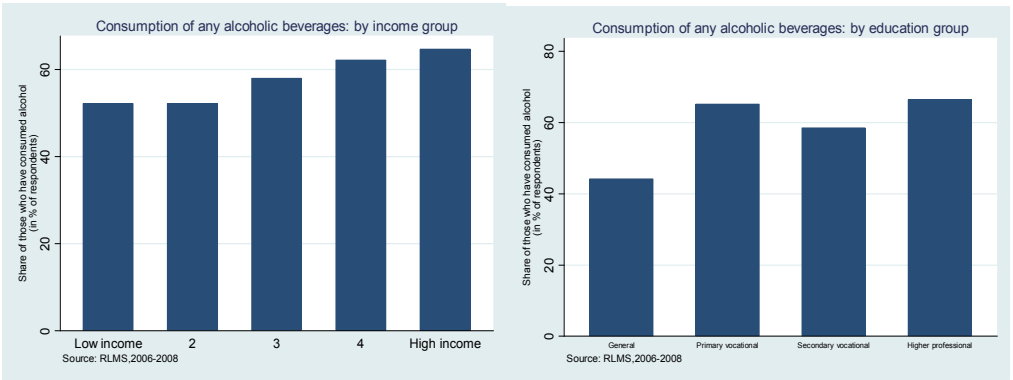


Fig. 8. Consumption of alcohol in Russia: by income group (left) and educational group (right). Percentage of total answers.

The share of abstainers (those who do not drink any alcohol) in the 18-24 and 14-16 age groups is shown in Fig.9. Among teenagers aged 14-16 about 25% consume alcohol and 75% are abstainers. The share of those who do not consume alcohol has grown somewhat in recent years. A similar trend of growing abstinence is revealed in the 18-24 age group: among women that share increased from 23% in 2006 to almost 30% in 2008; among men the share of non-drinkers in the 18-24 age group increased from 14.4% to 20.4%.

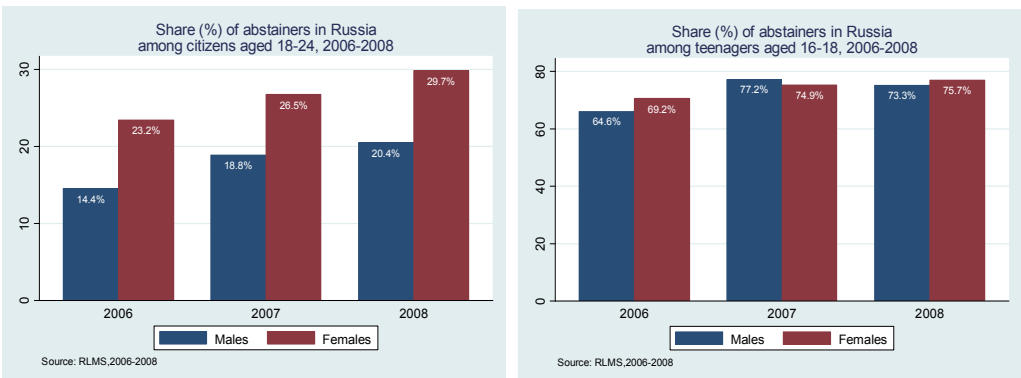


Fig. 9. Share of abstainers (non-drinkers) in Russia: 18-24 age group (left) and 14-16 age group (right). Percentage of total number of respondents in age groups.

The share of drinkers in different age groups is an important but not the only characteristic of a nation's alcohol consumption. Thus, in France, which has managed to diminish total alcohol consumption in recent years, more than 90% of the adult population and nearly 80% of persons aged 17-19 consume alcohol (WHO Global Status Report on Alcohol, 2004).

However, the fact that the overwhelming majority drink dry wine, and then in small or medium doses, puts France in an upward trend in terms of life expectancy. Russia is still characterized by the predominant consumption on strong alcoholic beverages.

The structure of alcohol consumption in Russia as reflected in the share of those who consume this or that type of drink, is shown in Fig.10. As seen from the charts, in Russia about 70% of men and nearly 50% of women drink beer. The next most popular drink is vodka and other strong spirits, consumed by more than 60% of men and 37% of women. About 12% of men and 5% of women drink home-made alcohol. About 40% of women and 11% of men drink dry wine or champagne. Thus, the most popular alcoholic beverages in Russia are vodka and home-produced alcohol, on the one hand, and beer on the other. At the same time, if one recalls Fig.2 which shows total consumption of beer and spirits in Russia in litres of pure alcohol, the absolute predominance of vodka cannot be disputed, whereas beer and wine account for only a small share of total alcohol consumption. This indicates the type of consumption: vodka is drunk more frequently and in larger quantities whereas beer and wine is drunk less frequently and in smaller quantities.

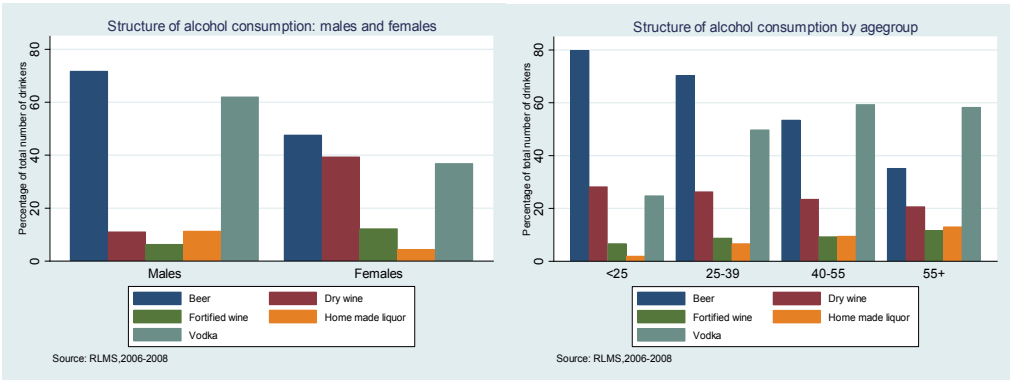


Fig. 10. Structure of alcohol consumption in Russia: by gender (left) and by age groups (right). Percentage of total number of drinkers.

It is notable that the consumption of beer, on the one hand, and of vodka and home brewed alcohol on the other reveals substantial differences by age group (Fig.10, right-hand side). Indeed, the distribution of beer consumers is tilted towards younger age groups whereas consumption of vodka is more characteristic of older age groups. This picture may attest to the beginnings of change in the pattern of consumption and a shift from predominantly strong alcoholic beverages towards beer. Whether the trend turns out to be sustained remains to be seen.

The structures of alcohol consumption by income groups and education are shown in Fig.11. The share of vodka and beer in the structure of consumption in different income groups is approximately the same, with beer consumption slightly higher in the lower-income groups. At the same time there is a marked trend of increased share of those who drink wine and champagne in the higher-income groups. Thus in the first (lowest) quantile only 16% of alcohol consumers drink wine and in the fifth (top) quantile the percentage is 31%. Similarly the consumption of wine and champagne is more common among people with a higher level of education. The lower share of vodka and beer drinkers in the groups with a higher

education may indicate a replacement of these drinks with wine among the better educated groups.

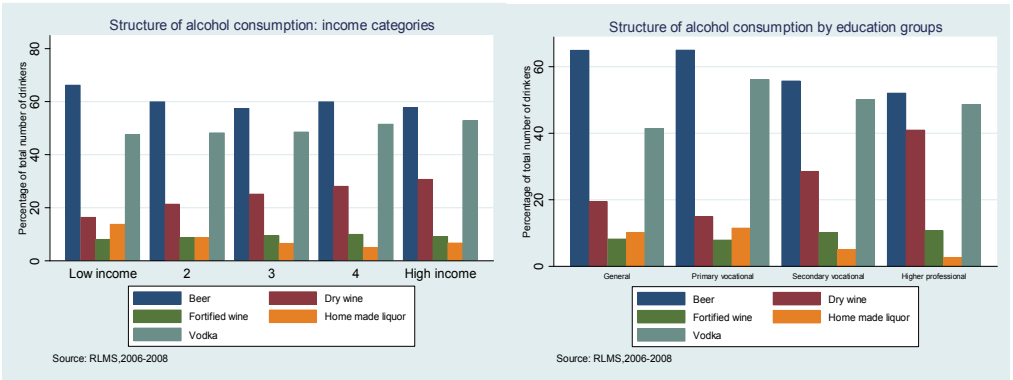


Fig. 11. Structure of alcohol consumption in Russia: by income group (left) and by educational group (right). Percentage of total number of drinkers.

A comparison of the frequency of alcohol consumption in Russia and some West European countries (Table 1) shows a similarity with Finland and Sweden. Indeed, the percentage of men who frequently consume alcohol (every day or 4-5 times a week) is 6.3% in Russia, 8% in Finland and 7% in Sweden. A further 35% of Russian men drink once or 2-3 times a week. Similar figures are reported in Sweden (40%) and Finland (50%). Russian women do not drink more frequently than women in the Nordic countries. This type of consumption contrasts with consumption in the southern countries: in Italy, for example, 45% of men and 30% of women drink (usually wine) every day or 4-5 times a week. The figures for France are 26% for men and 11% for women.

	Every day	4-5 times a week	2-3 times a week	Once a week	2-3 times a month	Once a month	Once or several times a year	Do not drink
Males:								
Finland	4	4	20	32	19	7	8	6
France	21	5	19	23	7	5	6	13
Germany	12	6	24	18	11	11	7	12
Italy	42	3	17	14	4	4	6	11
Sweden	3	4	16	24	23	12	12	7
UK	9	16	31	18	8	4	4	11
Russia (RLMS)	2.8	3.5	15.4	19.6	20.3	8.4	n.a.	30
Females:								
Finland	2	2	7	22	22	14	24	8
France	9	3	10	16	9	12	14	27
Germany	5	2	13	20	15	10	17	18
Italy	26	4	10	12	8	4	14	22
Sweden	1	1	5	17	2	17	23	13
UK	5	6	18	22	12	10	11	14
Russia (RLMS)	0.5	0.5	3.5	10	20.5	15	n.a.	50

Source: Alcohol in post-war Europe (2001), Table 5.1, p. 107 for Western Europe and authors' calculations for Russia

Table 1. Frequency (%) of alcohol consumption, West Europe and Russia, men and women

The frequency of alcohol consumption hardly varies in different income groups and varies only slightly by education group (Fig.12). The better educated drink less frequently: the share of those who drink every day or 4-6 times a week among graduates of vocational training schools is 7% and among graduates of secondary professional schools and higher education institutions 4%. The number of those who drink 2-3 times a month is 5% higher among university graduates: 37% versus 32% among graduates of vocational secondary schools. These differences reflect more moderate alcohol consumption among women who happen to be better educated.

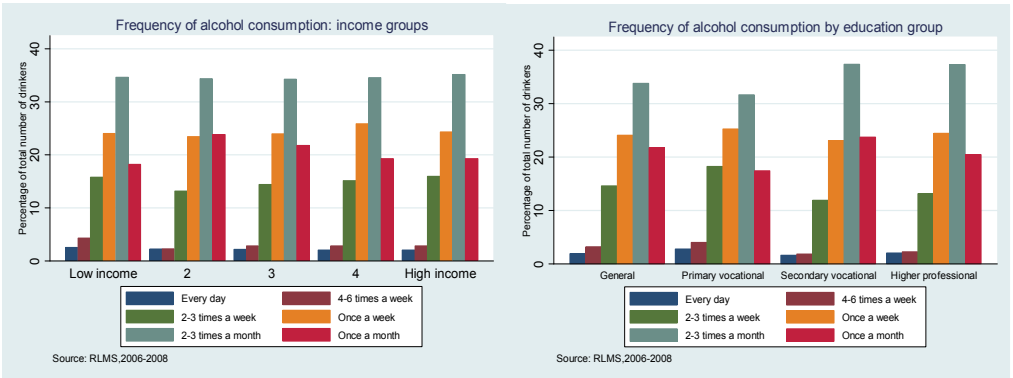
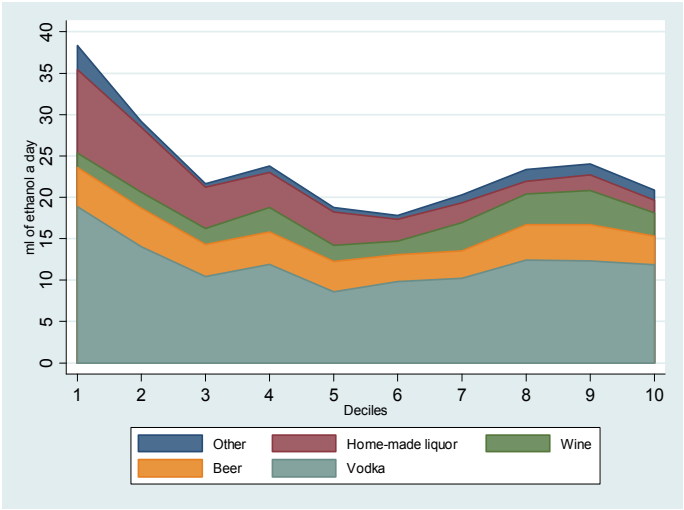


Fig. 12. Frequency of alcohol consumption in Russia: by income group (left) and by education group (right). Percentage of total number of drinkers.

Differences in the structure of alcohol types and frequency of consumption manifest themselves in differences in the average quantity of ml of ethanol consumed. The differences in the average daily consumption of ethanol between different income groups are shown in Fig.13. As seen from the chart, consumption of ethanol is the highest in the first three income groups (the poorest) which is due to the prevalence of the consumption of vodka and home brewed alcohol in these groups. At the same time the high income groups in the 8th, 9th and 10th deciles show a high consumption of ethanol comparable to the 3rd and 4th income groups, which is a consequence of the high consumption of vodka in top deciles.

The structure of alcoholic drinks consumption by frequency is shown in Fig. 14. As seen from the chart, the prevalent type of alcohol consumption in Russia is the drinking of vodka separately or in combination with other drinks. This type of consumption is more pronounced among frequent drinkers: almost 60% of those who drink every day or 4-6 times a week drink vodka, separately or in combination with other drinks. Among those who drink less frequently vodka drinkers account for more than half. At the same time it is noticeable that the share of those who drink wine and beer, but do not drink vodka is higher among the groups of infrequent drinkers (once a year or 2-3 times a month) than among other groups.



Source: Authors calculations based on RLMS-HSE, 2005 using Andrienko and Nemtsov, 2005 approach.

Fig. 13. Structure of alcohol consumption by income groups in Russia (1 – the poorest, 10 – the richest), ml ethanol a day

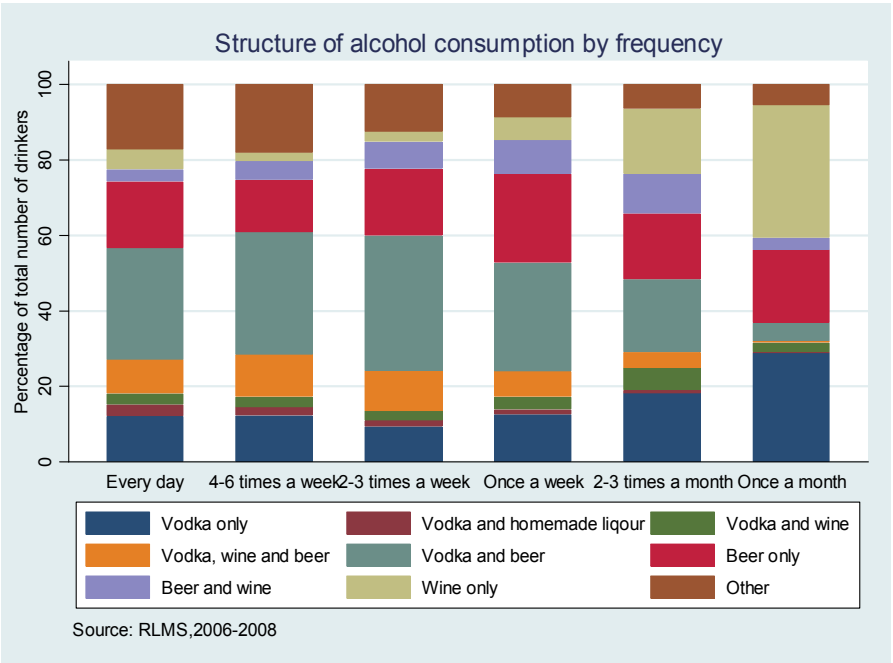


Fig. 14. Structure of alcohol consumption by frequency (% of those who drank alcohol).

4. The impact of alcohol consumption on health and risk of death

This section presents the results of the estimates of the impact of alcohol consumption on the health and risk of death in Russia based on the data of the Russian Longitudinal Monitoring

Survey. The RLMS-HSE questionnaire contains a wide range of questions characterizing various aspects of the behavior of individuals in the family and in the labor market and detailed information on harmful habits and health indicators. In addition, the longitudinal character of the study makes it possible to trace the behavior trajectories of respondents over many years and to study the causes of death.

The regressions cited below test the impact of frequent alcohol consumption on health (Tables 2 and 3) and the risk of death (Table 4) and separates the overall effect of frequent consumption of alcohol from frequent consumption of vodka (straight or in combination) and divide frequent alcohol consumption into frequent consumption of vodka (straight or mixed) and frequent consumption of beer (separately or mixed with drinks other than vodka). RLMS makes it possible to determine how frequently and what beverages an individual drinks. We have identified a group of those who consume alcohol every day or 4-6 times a week, calling it "frequent drinkers". In addition we have identified two subgroups among the frequent drinkers: those who drink vodka, separately or in combination with other drinks ("frequent vodka drinkers") and those who drink beer, but do not drink vodka ("frequent beer drinkers").

The health variable used is the person's own assessment of his/ her state of health (very bad, bad, satisfactory, good and very good). We use all the categories as well as define the binary variable: bad and very bad health versus all the other variants. Death is registered in the sample on the basis of the information provided by the household head when the unit is surveyed at least two rounds in a row. A household head is asked to report whether any household member is missing during the survey round and the reason for that member being not in the household. One of the reasons reported is the death of the household member. More details on the measurement and methodology could be found in Denisova (2010).

In all cases we control for gender, age, the respondent's education, per capita household income, place of residence, body mass index and smoking (whether or not the respondent smokes). We also control for the individual's assessment of his own social status on a nine-point scale ("respected – not respected"). This makes it possible to take into account the impact of constant psychological stress on a person's health and separate that impact from the impact from alcohol consumption.

The impact of alcohol consumption on health is estimated based on pooled cross-section for 1994-2007 (Table 2) and on a panel for the same years (Table 3). The impact of alcohol consumption on the risk of death is estimated with Gompertz proportional hazard model (Table 4).

The results of the estimates of the impact of alcohol consumption on health in Table 2 show that frequent alcohol consumption harms health. Thus, frequent alcohol consumption increases the probability of having bad or very bad health by 7 percentage points. Moreover, frequent alcohol consumption that includes vodka leads to health deterioration (the risk of bad health increases by 9 percentage points) whereas frequent consumption of beer does not have a statistically significant effect on health. The negative impact of frequent alcohol consumption on health is stable regardless of the method of assessment and of control for individual specific recorded effects in particular (Table 3). Frequent alcohol consumption increases the probability of bad or very bad health by 17 percentage points, with the entire effect caused by frequent consumption of vodka.

	Self-accessed health						Bad and very bad health		
	Males and females		Males		Females		Males and females	Males	Females
Age	-0.02	-0.02	-0.021	-0.021	-0.018	-0.018	0.035	0.037	0.032
	[0.000]***	[0.000]***	[0.000]***	[0.000]***	[0.000]***	[0.000]***	[0.001]***	[0.001]***	[0.001]***
Gender: Males	0.209	0.209					-0.202		
	[0.006]***	[0.006]***					[0.017]***		
Married	0.032	0.032	0.031	0.031	0.029	0.029	-0.14	-0.172	-0.12
	[0.005]***	[0.005]***	[0.009]***	[0.009]***	[0.007]***	[0.007]***	[0.015]***	[0.026]***	[0.019]***
Junior or secondary professional education	-0.01	-0.01	-0.029	-0.029	0.004	0.004	-0.085	-0.063	-0.099
	[0.006]*	[0.006]*	[0.009]***	[0.009]***	[0.007]	[0.007]	[0.016]***	[0.026]**	[0.020]***
Higher education	0.063	0.063	0.056	0.056	0.057	0.056	-0.211	-0.219	-0.193
	[0.008]***	[0.008]***	[0.012]***	[0.012]***	[0.009]***	[0.009]***	[0.022]***	[0.038]***	[0.027]***
Log of real per capita income, 1992 prices	0.022	0.022	0.019	0.019	0.023	0.023	-0.075	-0.088	-0.066
	[0.003]***	[0.003]***	[0.004]***	[0.004]***	[0.004]***	[0.004]***	[0.008]***	[0.012]***	[0.010]***
Self-perceived status, respect rank on 9-step ladder	0.027	0.027	0.035	0.036	0.018	0.018	-0.035	-0.053	-0.02
	[0.001]***	[0.001]***	[0.002]***	[0.002]***	[0.002]***	[0.002]***	[0.004]***	[0.006]***	[0.005]***
Frequent alcohol drinker	-0.025		-0.016		-0.021				
	[0.014]*		[0.015]		[0.033]				
Frequent vodka drinker (pure and in mix)		-0.029		-0.023		-0.025	0.092	0.084	0.066
		[0.017]*		[0.018]		[0.043]	[0.043]**	[0.048]*	[0.113]
Frequent beer drinker (no vodka)		0.002		0.012		0.021	-0.105	-0.122	-0.187
		[0.031]		[0.035]		[0.066]	[0.101]	[0.115]	[0.212]
Smokes	-0.06	-0.06	-0.041	-0.041	-0.07	-0.071	0.038	-0.024	0.07
	[0.006]***	[0.006]***	[0.007]***	[0.007]***	[0.008]***	[0.008]***	[0.017]**	[0.024]	[0.026]***
Body mass index	0.029	0.029	0.111	0.111	0.007	0.007	-0.064	-0.184	-0.017
	[0.017]*	[0.017]*	[0.025]***	[0.025]***	[0.011]	[0.011]	[0.003]***	[0.058]***	[0.028]
Body mass index squared/1000	-0.538	-0.539	-1.891	-1.891	-0.241	-0.241	1.158	2.998	0.503
	[0.304]*	[0.304]*	[0.468]***	[0.468]***	[0.195]	[0.195]	[0.047]***	[1.082]***	[0.452]
Unemployed	-0.002	-0.002	0.004	0.004	-0.002	-0.002	0.082	0.031	0.1
	[0.011]	[0.011]	[0.017]	[0.017]	[0.015]	[0.015]	[0.032]**	[0.052]	[0.042]**
Urban settlement	-0.061	-0.061	-0.062	-0.062	-0.06	-0.06	0.034	0.046	0.026
	[0.005]***	[0.005]***	[0.008]***	[0.008]***	[0.007]***	[0.007]***	[0.015]**	[0.023]**	[0.019]
Constant	3.197	3.196	2.288	2.287	3.575	3.575	-0.87	0.737	-1.705
	[0.230]***	[0.230]***	[0.321]***	[0.321]***	[0.155]***	[0.155]***	[0.084]***	[0.756]	[0.380]***
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	75037	75037	34083	34083	40954	40954	75265	34204	41061
R squared	0.19	0.19	0.17	0.17	0.18	0.18	0.12	0.12	0.12

Standard errors in brackets; * significant at 10%; ** significant at 5%; *** significant at 1%

Table 2. Influence of alcohol consumption on health, pooled cross-section, 1994-2007

	Self-accessed health						Bad and very bad health			
	Males and females		Males		Females		Males and females		Males	Females
Age	-0.016	-0.016	-0.023	-0.023	-0.01	-0.01	0.004	0.004	0.006	0.003
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.000]***	[0.000]***	[0.000]***	[0.000]***
Married	0.031	0.031	0.038	0.038	0.034	0.034	-0.002	-0.002	-0.006	-0.001
	[0.007]***	[0.007]***	[0.011]***	[0.011]***	[0.008]***	[0.008]***	[0.003]	[0.003]	[0.005]	[0.005]
Log of real per capita income, 1992 prices	0.007	0.007	0.006	0.006	0.008	0.008	-0.005	-0.005	-0.005	-0.006
	[0.003]**	[0.003]**	[0.004]	[0.004]	[0.004]**	[0.004]**	[0.001]***	[0.001]***	[0.002]**	[0.002]***
Self-perceived status, respect rank on 9-step ladder	0.01	0.01	0.013	0.013	0.007	0.007	-0.002	-0.002	-0.003	-0.002
	[0.001]***	[0.001]***	[0.002]***	[0.002]***	[0.002]***	[0.002]***	[0.001]***	[0.001]***	[0.001]***	[0.001]**
Frequent alcohol drinker	-0.028		-0.008		-0.098		0.017			
	[0.013]**		[0.015]		[0.031]***		[0.006]***			
Frequent vodka drinker (pure and in mix)		-0.028		-0.008		-0.13		0.017	0.01	0.056
		[0.015]*		[0.017]		[0.039]***		[0.007]**	[0.007]	[0.021]***
Frequent beer drinker (no vodka)		0.017		0.042		-0.037		-0.005	-0.004	-0.015
		[0.029]		[0.035]		[0.059]		[0.014]	[0.015]	[0.032]
Smokes	-0.011	-0.012	0.01	0.01	-0.048	-0.048	-0.012	-0.012	-0.022	0.003
	[0.010]	[0.010]	[0.013]	[0.013]	[0.014]***	[0.014]***	[0.005]***	[0.005]**	[0.006]***	[0.008]
Body mass index	0.015	0.015	0.052	0.052	0.005	0.005	-0.008	-0.008	-0.021	-0.005
	[0.003]***	[0.003]***	[0.008]***	[0.008]***	[0.003]	[0.003]	[0.001]***	[0.001]***	[0.003]***	[0.002]***
Body mass index squared/1000	-0.15	-0.15	-0.71	-0.711	-0.057	-0.057	0.073	0.073	0.277	0.045
	[0.036]***	[0.036]***	[0.130]***	[0.130]***	[0.037]	[0.037]	[0.018]***	[0.018]***	[0.056]***	[0.020]**
Constant	3.451	3.451	3.252	3.254	3.338	3.338	0.15	0.149	0.262	0.163
	[0.050]***	[0.050]***	[0.117]***	[0.117]***	[0.060]***	[0.060]***	[0.025]***	[0.025]***	[0.050]***	[0.033]***
Number of observations	75054	75054	34097	34097	40957	40957	75283	75283	34219	41064
Number of individuals (groups)	19026	19026	9025	9025	10002	10002	19056	19056	9043	10014
R squared	0.16	0.16	0.15	0.15	0.16	0.16	0.06	0.06	0.06	0.05

Standard errors in brackets; * significant at 10%; ** significant at 5%; *** significant at 1%

Table 3. Influence of alcohol consumption on health, panel fixed effects, 1994-2007

The results of the assessment of the impact of alcohol on risk of death are shown in Table 4. As seen from the table, frequent consumption of alcohol, above all of strong spirits, increases the risk of death. Simple consumption of vodka or beer does not yield a statistically significant effect. Frequent consumption of alcohol increases the risk of death by 60 percentage points. Frequent consumption of vodka increases the risk of death by 66 percentage points whereas frequent consumption of beer does not have a statistically significant effect.

	(1)	(2)	(3)	(4)
Economic well-being				
Household in poverty: the 1st poverty episode	0.86	0.853	0.86	0.854
	[0.045]***	[0.045]***	[0.045]***	[0.045]***
Household in poverty: the 2nd, 3d, ... poverty episodes	1.373	1.338	1.369	1.343
	[0.227]*	[0.230]*	[0.226]*	[0.232]*
Consumption decile (within year)	0.981	0.98	0.98	0.975
	[0.154]	[0.155]	[0.155]	[0.154]
Self-perceived status				
Economic rank on 9-step ladder	0.973	0.97	0.972	0.971
	[0.035]	[0.035]	[0.035]	[0.035]
Respect rank on 9-step ladder	0.947	0.945	0.946	0.947
	[0.023]**	[0.023]**	[0.023]**	[0.023]**
Stress indicator				
Concern about getting necessities	1.088	1.065	1.083	1.077
	[0.113]	[0.112]	[0.112]	[0.113]
Habits				
Smokes	1.582	1.577	1.584	1.563
	[0.193]***	[0.191]***	[0.193]***	[0.188]***
Frequent alcohol drinker	1.594			1.514
	[0.282]***			[0.273]**
Vodka/hard liquids drinker		1.142		1.117
		[0.123]		[0.124]
Beer drinker		1.021		0.999
		[0.136]		[0.132]
Frequent vodka drinker			1.663	
			[0.324]***	
Frequent beer drinker			1.243	
			[0.726]	
Alchohol availability				
Relative price of vodka to bread in locality	1.015	1.015	1.015	1.015
	[0.011]	[0.011]	[0.011]	[0.011]
	(1)	(2)	(3)	(4)
Labor market experience				
Unemployed	1.495	1.503	1.498	1.498
	[0.363]*	[0.361]*	[0.365]*	[0.363]*
Experience as entrepreneur/self-employed	0.472	0.472	0.471	0.472
	[0.182]*	[0.182]*	[0.181]*	[0.182]*
Mobile in labor market	0.488	0.493	0.489	0.492
	[0.087]***	[0.087]***	[0.087]***	[0.087]***
Health care accessibility				
Could not afford or find prescribed medicine	1.179	1.144	1.174	1.149
	[0.264]	[0.269]	[0.262]	[0.270]
Gender: Males	3.478	3.484	3.481	3.434
	[0.453]***	[0.476]***	[0.453]***	[0.466]***
Social and individual human capital				
Married	1.081	1.07	1.077	1.073
	[0.114]	[0.114]	[0.114]	[0.115]
Family size, number of people in family	1.161	1.163	1.161	1.163
	[0.041]***	[0.042]***	[0.041]***	[0.042]***
Children in family	0.769	0.754	0.769	0.753
	[0.112]*	[0.110]*	[0.112]*	[0.111]*
Education: secondary school and below - reference category				
Junior or secondary professional	0.847	0.836	0.852	0.839
	[0.087]	[0.087]*	[0.089]	[0.087]*
University degree or higher	0.649	0.647	0.652	0.649
	[0.123]**	[0.124]**	[0.124]**	[0.124]**
Urban settlement	0.758	0.744	0.755	0.743
	[0.072]***	[0.076]***	[0.071]***	[0.076]***
Health undicators	Yes***		Yes***	
Gompertz function coefficients	0.053	0.053	0.053	0.053
	[0.004]***	[0.004]***	[0.004]***	[0.004]***
Observations	70715	70513	70715	70513
No. of subjects	17606	17596	17606	17596
No. of failures	420	418	420	418
Log Pseudolikelihood	-603.07	-602.64	-603.24	-600.36

Robust standard errors in brackets; * significant at 10%; ** significant at 5%; *** significant at 1%

Table 4. Determinants of mortality, working age population, 18-65, parametric Gompertz regression

The results obtained can be represented as differences in survival functions for those who frequently drink strong alcoholic beverages and those who rarely or never consume strong liquor. Such functions are represented in Fig. 15. As seen from the ratio of the curves, frequent consumption of vodka shortens life by an average 9-10 years. At the same time, as noted above, frequent beer consumption has not yielded a statistically significant result.

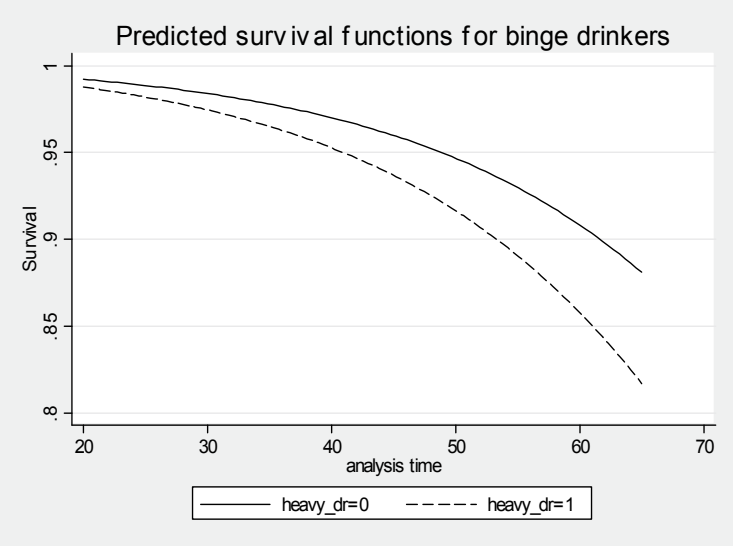


Fig. 15. Predicted survival curves: frequent drinkers of strong liquor (broken line) and infrequent drinkers and non-drinkers (solid line). The forecast is based on estimates presented in Table 3.

Thus, the regression analysis based on longitudinal data makes it possible to isolate the impact of various types of alcohol consumption on the health and mortality controlled for the impact of other groups of factors. Our results attest to a strong negative impact of frequent alcohol consumption on health and the risk of death. Moderate alcohol consumption does not exert a statistically significant effect. In addition, the negative impact of frequent consumption of strong alcoholic beverages is greater than the effect of frequent consumption of wine and beer: frequent consumption of vodka shortens life by an average 9-10 years, whereas no statistically significant impact of frequent beer drinking has been revealed.

5. Anti-alcohol policy measures: The experience of European countries

The experience of the countries of Northern Europe in encouraging people to switch from the consumption of mainly strong alcoholic beverages to lighter ones (beer and wine) shows active use of excise policy measures. Table 5 contains information on excise rates on alcohol, intermediate products, wine and beer in European countries in 2010. As seen from the table, the excise rates in the North European countries are significantly higher than in other countries, and excise on alcohol is 2-4 times higher than excise on beer (in liters of pure alcohol). In some other European countries the gap between excise on spirits and beer is six-fold, although excise is much lower than in the North European countries in absolute terms.

Similar calculations of the ratio of excise on various types of alcoholic drinks in Russia could be found in the bottom line in Table 5. As seen from the table, excise on all alcoholic beverages in Russia is several times lower than in European countries. Moreover, the structure of excise tax in Russia differs drastically from the structure in excise in Europe. Russia has a very low excise on distilled spirits: 5.25 euros against 49.21 euros in Sweden (and even 85.36 euros in Norway) and 10 euros in Austria. The excise rate on wine in Russia is close to the rates of wine-consuming countries of southern Europe, a mere 80 eurocents. The rate of tax on beer increased significantly in 2010 but is much lower than the similar rates in North Europe, UK and Ireland in absolute terms. It is though comparable – and even higher if income differences are taken into account – with excise rates in beer-consuming countries. Beer excise amounts to 26 euros in Finland, 19.87 euros in Ireland, 18.08 euro in UK, 17.07 euros in Sweden, but is only 1.73 euros in Germany, 4.4 euros in Austria, 4.9 euros in Netherlands and 4.5 euros in Russia. In addition, strong beer (upward of 8.6%) is common in Russia, unlike in Europe.

Country	*Distilled spirits*	*Wine (11%)*	*Beer (5%)*
Austria	10.03	0	4.4
Belgium	17.52	4.28	3.76
Denmark	20.14	7.42	6.84
Finland	39.4	25.47	26
France	15.12	0.3	2.71
Germany	13.04	0	1.73
Greece	15.7	0	3.58
Ireland	31.13	23.81	19.87
Italy	8	0	5.17
Netherlands	15.04	6.23	4.9
Portugal	10	0	2.76
Spain	8.3	0	2
Sweden	49.21	19.25	17.07
UK	24.85	21.59	18.08
Russia	5.25	0.8	4.5

Source: CEPS, Summary of EU Member States at http://www.europeanspirits.org/OurIndustry/TaxationIndustry.asp (Rates as of January 2010) for Europe and Federal Law No. 171-FZ for Russia (converted according to 40 rubles per euro exchange rate).

Table 5. Alcohol excise rates (euros per litre of pure alcohol), Russia and Europe, 2010

On the whole excise rates on alcohol in Russia are rather low, especially for distilled spirits, even with due account of the differences in the purchasing power of the population in Europe and Russia. Excise rates on distilled spirits in Russia are strikingly low when compared to the rates in countries with unhealthy high alcohol consumption and high share of consumption of hard liquids. Moreover, the ratios of excise rates on distilled spirits and beer are in sharp contrast to those in all European countries. Wine excise rate is comparable to wine-consuming countries of Europe. The policy of reducing alcohol consumption dictates an increase of excise on all alcoholic products. If Russians are to be induced to consume less alcohol and strong alcoholic beverages in particular, excise on distilled spirits must grow faster than on other goods.

The effectiveness of the use of price mechanisms to limit alcohol consumption depends on how price-elastic demand for alcoholic beverages is. Assessments of the elasticity of demand for alcoholic beverages in the European countries and Russia will be found in Table 6.

Country	*Price elasticity*	*Income elasticity*
Finland, Sweden, Norway	-0.782	0.752
Austria, France, Greece, Italy, Portugal and Spain	-0.216	0.752
Belgium, Denmark, Ireland, UK	-0.495	0.752
Netherlands	-1.466	0.752
Russia		
Vodka	-1.774	0.524
Beer	-3.017	1.114
Wine	-1.045	1.304

Source: Leppanen et al, 2001 for European countries, Andrienko and Nemtsov, 2005 for Russia

Table 6. Elasticity of demand for alcoholic beverages in the countries of Europe and Russia

As seen from the table the assessment of the elasticity of demand for vodka, beer and wine depending on price in Russia is much higher than the same indicators for the European countries and is comparable to the elasticity of demand in the Netherlands. This suggests that price measures of influencing alcohol consumption – increasing excise on strong beverages in order to encourage people to switch to types of alcohol consumption that cause less harm to health – can be effective. In Russia the experience of the past decades has revealed a relative cheapening of vodka. Thus in the late 1980s the ratio of vodka and beer prices was such that one could buy nine bottles of beer for the price of one bottle of vodka. At present it can buy on average 3-4 bottles of beer (and some vodkas are even cheaper).

Along with price measures, an effective anti-alcohol policy must include other measures: the system of control of the production and sale of alcohol, restriction of the sale of alcohol outside restaurants, bars, etc., restrictions on the minimum age of the customer to whom alcohol can be sold, restrictions on alcohol advertising, on sponsorship of sporting and youth events, and on consumption of alcohol in public places. All the European countries, and not only the northern ones, have significantly strengthened their anti-alcohol policies in the last 50 years.

6. Conclusions

Mortality dynamics in Russia are due in large part to excessive alcohol consumption. In Russia consumption of strong alcoholic beverages exceeds the consumption of beer and wine both in terms of the aggregate volume and prevalence among the population. This "northern" type of consumption is characteristic of all the income and education groups.

There is a trend for vodka consumption to shift towards the older age group while young people are switching to the consumption of beer. Whether this is a sustained trend is unclear. There is a trend for better educated and wealthier people to switch to the consumption of wine rather than hard liquor.

An analysis of the impact on health and mortality attests to a strong negative impact of frequent alcohol consumption. Moderate alcohol consumption does not produce a statistically significant negative effect.

The negative impact of frequent consumption of hard liquor exceeds that of the consumption of wine and beer. Frequent consumption of vodka shortens the lifespan by an average 9-10 years while no statistically significant impact has been revealed of frequent consumption of beer.

An active anti-alcohol policy must include both price and non-price measures. The policy aimed at reducing alcohol consumption calls for a rise of excise on all the alcoholic products. If Russians are to be encouraged to switch to lighter drinks, excise on hard liquor must grow faster than other excise rates.

There is a large untapped potential for the use of non-price anti-alcohol measures aimed at reducing alcohol consumption in Russia. The experience of North European countries which have succeeded in switching from the consumption of predominantly hard liquor to the consumption of beer and wine and significantly cutting the consumption of hard liquor is of particular relevance.

7. References

[1] *Alcohol in Postwar Europe.* Consumption, Drinking Patterns? Consequences and policy responses in 15 European countries (2001) Thor Norstrom (editor). National Institute of Public Health, Sweden.

[2] Alcohol per capita consumption, patterns of drinking and abstention worldwide after 1995. Appendix 2. *European Addiction Research*, 2001, 7(3):155–157.

[3] Andrienko, Yuri, and Alexander Nemtsov (2005). "Estimation of Individual Alcohol Demand," *Economics Education and Research Consortium, Working Paper series*, 05/10.

[4] Babor, T. F., Caetano, R., Caswell, S., Edwards, G., Giesbrecht, N., Graham, K., et al. (2003). *Alcohol: No ordinary commodity. Research and public policy*. Oxford, United Kingdom: Oxford University Press.

[5] Braninerd, Elizabeth and David M.Cutler (2005). "Autopsy of an Empire: Understanding Mortality in Russia and the Former Soviet Union", *Journal of Economic Perspectives,* 19, 1, pp.107-130.

[6] Bruun, K., Edwards, G., Lumio, M., Makela, K., Pan, L., Popham, R. E. et al. (1975) *Alcohol Control Policies in Public Health Perspective.* Finnish Foundation for Alcohol Studies, Helsinki.

[7] Chaloupka, F., Grossman, M. and Saffer, H. (2002). "The effects of price on alcohol consumption and alcohol-related problems". *Alcohol Research and Health.* (26)1: 22-34.

[8] Denisova, Irina (2010) "Adult mortality in Russia: a microanalysis", *Economics of Transition,* Vol.18(2), 333-363.

[9] Di Castelnuovo, Augusto, Serenella Rotondo, Licia Iacoviello, Maria Benedetta Donati, Giovanni de Gaetano (2002) "Meta-Analysis of Wine and Beer Consumption in Relation to Vascular Risk" *Circulation*, 105, 2836-2844.

[10] Edwards G. et al. (1994). *Alcohol policy and the public good.* Oxford: Oxford University Press (Издание на русском языке: Алкогольная политика и общественное

благо / Ред. Г. Эдвардс. Региональные публикации ВОЗ. Европейская серия № 80. 1998).

[11] Edwards, G., Anderson, P., Babor, T. F., Casswell, S., Ferrence, R., Giesbrecht, N., Godfrey, C., Holder, H.D., Lemmens, P., Mäkelä, K., Midanik, L., Norström, T., Österberg, E., Romelsjö, A., Room, R., Simpura, J. & Skog, O.-J. (1994) *Alcohol Policy and the Public Good* (Oxford, Oxford University Press).

[12] European Health for All database (HFA-DB). 2008. Copenhagen, WHO Regional Office for Europe (http://www.euro.who.int/hfadb).

[13] Foxcroft, D. R., Ireland, D., Lister-Sharp, D. J., Lowe, G., and Breen, R. (2003). "Longer-term primary prevention for alcohol misuse in young people: A systematic review". *Addiction, 98*, 397–411.

[14] Gilinskiy Y. (2000) Analysis of statistics on some forms of social deviation in St. Petersburg from 1980 to 1995. In: Leifman H, Edgren-Henrichson N, eds. *Statistics on alcohol, drugs and crime in the Baltic Sea region.* Helsinki, Nordic Council for Alcohol and Drug Research (NAD)

[15] Grube, J. W. & Nygaard, P. (2001). "Adolescent Drinking and Alcohol Policy." *Contemporary Drug Problems*, 28, 87-131.

[16] Harkin, A.M., Anderson, P. & Lehto, J. (1995). *Alcohol in Europe: A Health Perspective*. Copenhagen: World Health Organization Regional Office for Europe.

[17] Karlsson, T. and Österberg, E. (2001). "*A Scale of Formal Alcohol Control Policy in 15 European Countries.*" Nordisk Alkohol& Narkotikatidskrift (Nordic Studies on Alcohol and Drugs), (EnglishSupplement): 117-31.

[18] Leon, David (2007). "Hazardous Alcohol Drinking and Premature Mortality in Russia: a Population Based Case-Control Study," *Lancet,*Vol.369, Issue 9578, pp.2002-2009.

[19] Leon, David A., L. Chenet, Vladimir Shkolnikov, Sergei Zakharov, Judith Shapiro, Galina Rakhmanova, Sergei Vassin, and Martin McKee (1997). "Huge Variation in Russian Mortality Rates 1984-94: Artifact, Alcohol, or What?" *Lancet*, 350, pp.383-88.

[20] Leppanen, K., Sullstrom, R. & Suoniemi, I. (2001) "*The Consumption of Alcohol in Fourteen European Countries. A Comparative Econometric Analysis*" (Helsinki, Stakes).

[21] Makela P., K. Tryggvesson, and I. Rossow (2002). "Who drinks more or less when policies change? The evidence from 50 years of Nordic studies. The effects of Nordic alcohol policies: Analyses of changes in control systems". Ed. by R. Room, pp. 17-70. Helsinki, Nordic Council for Alcohol and Drug Research.

[22] Makela K., E. Osterberg, and P. Sulkunen (1981). "Drinking in Finland. Increasing alcohol availability in a monopoly state. Alcohol, society, and the state 2: The history of control policy in seven countries". Ed. by E. Single, P. Morgan, and J. de Lint, pp. 31-60. Toronto: Addiction Research Foundation.

[23] Nemtsov A. (2002). "Alcohol-related harm losses in Russia in the 1980s and 1990s". *Addiction.* 97. 1413—1425.

[24] Osterberg E. (1995). "Do alcohol prices affect consumption and related problems? Alcohol and public policy. Evidence and issues." Ed. by H. Holder and G. Edwards, pp. 145-163. Oxford: Oxford University Press

[25] Osterberg E. (2001). *Pricing and Taxation.* Handbook on alcohol dependence and related problems / Ed. by N. Heather, T. Peters, T. Stockwell, pp. 685-698. London: Wiley.

[26] Österberg, E. & Karlsson, T. (2002) *Alcohol Policies in EU Member States and Norway. A Collection of Country Reports* (Helsinki, Stakes).

[27] Rehn, N., Room, R., & Edwards, G. (2001). *Alcohol in the European Region - Consumption, Harm, and Policies.* Copenhagen: World Health Organization Regional Office for Europe.

[28] Shkolnikov, V.M., G.A Cornia, D.A. Leon, and F. Mesle (1998). "Causes of the Russian Mortality Crisis: Evidence and Interpretations," *World Development*, 26, 11, pp.1995-2011.

[29] Wagenaar A. and Toomey T. (2000). "Alcohol policy: gaps between current research." *Contemporary drug problems*, 27:681-733.

[30] Wagenaar, A. and Holder, H. (1995) "Changes in Alcohol Consumption Resulting from the Elimination of Retail Wine Monopolies: Result from Five U.S. States". *Journal of Studies on Alcohol*, Vol .56, No. 5.

[31] WHO Global Status Report on Alcohol 2004, Country Profiles, World Health Organization 2004

[32] WHO report on alcohol consumption, 2007

[33] Nemtsov A. (2009) Alcohol History of Russia: the latest period. Moscow: Librokom (in Russian)

19

Insomnia and Its Correlates: Current Concepts, Epidemiology, Pathophysiology and Future Remarks

Yuichiro Abe[1,2] and Anne Germain[3]
[1]Department of Psychophysiology,
National Instutute of Mental Health, NCNP, Tokyo,
[2]Policlinique, ASM13, Gentilly,
[3]Department of Psychiatry, University of Pittsburgh,
School of Medicine, Pittsburgh, Pennsylvania
[1]Japan
[2]France
[3]USA

1. Introduction

Insomnia is a common sleep disorder. People suffering from insomnia generally report not only sleep-related symptoms such as difficulty initiating, maintaining, obtaining sufficient restorative sleep, but also experience various daytime impairment reflective of sleep deficits (Buysse, 2008; Riemann et al., 2011). The generic term "insomnia" as a diagnostic entity is defined as a complaint of sleep problems coupled with impairment of daytime functioning, including reduced alertness, fatigue, exhaustion, dysphoria and other symptoms. The complaints have to endure for at least 4 weeks to be diagnosed as insomnia, according to the current diagnostic classification manual (Abe & Mishima, 2008).

Chronic insomnia is a "24-hour disease", meaning not only reduces the quality of sleep during the night, but also causes a variety of impairments in mental and physical functioning during the daytime (Bonnet & Arand, 1995, 2011). Although some patients who have this problem may not report it as such, inadequate sleep has been associated with reduced physical health and mental health (Morin & Espie, 2004; LeBlanc et al., 2007). Thus, many people are likely those who are in the "pre-insomnia" moment, and do not even consider themselves insomniacs (Bastien et al., 2004). Chronic insomnia is also associated with both human and socioeconomic costs, such as increased long-term absenteeism at work, reduced performance and productivity, and increased industrial accidents and health-care costs. This impact could be explained by three points: 1) comorbid mental (psychiatric) conditions, 2) comorbid medical conditions and 3) socioeconomic impact of insomnia (Mai & Buysse, 2008).

In primary care, practitioners usually prescribe medication such as hypnotics without for such insomnia complaints. However, the use of these sedative agents is often problematic, especially when patients have kept a good QOL activity in daily life (Riemann et al., 2011).

The mere augmentation of medication runs a risk of exacerbating daytime impairment itself. The continued widespread use of sedative medication to treat insomnia raises concern about the potential for long term tolerance and addiction, particularly where insomnia is the presenting complaint of missed diagnoses such as comorbid depression and anxiety disorder, or when adverse effects might be a problem—for example, falls in older adults (Riemann et al., 2011).

We will review about insomnia in terms of several aspects: its concept, epidemiology, pathophysiology, psychobehavioral correlates and possible psychiatric interventions. At the same time, we will show our own epidemiological study about Japanese people with insomnia based on the general population sample, and present some clinical case studies in order to describe several aspects of insomnia comorbid with mental disorders. We will also mention the correlates of nightmares, sleep disturbances related to suicidality and alcoholism as current important clinical and research topics. Finally, we will comment on future remarks based on the current society in Japan aftermath of Tsunami disaster in March 2011. In discussing mainly insomnia and nightmare, we used the terms "sleep disturbances" and "sleep problems" interchangeably in this paper, following the context.

2. Current definition and prevalence of insomnia

The reported prevalence of insomnia in the general population varies widely, ranging between 4.4% and 48%, depending on sample characteristics and the definition of insomnia (Ohayon, 2002). According to the American Sleep Disorders Association International Classification of Sleep Disorders (ICSD-2) published in 2005, its coding manual, insomnia refers to "a repeated difficulty with sleep initiation, duration, consolidation, or quality that occurs despite adequate time and opportunity for sleep and results in some form of daytime impairment and lasting for at least one month." (AASM, 2005).

2.1 DSM-5 proposed criteria of insomnia

The major current diagnostic systems ICD-10 (International Classification of Disorders 10th edition) and DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, APA) includes sections on insomnia and several sleep disorders. Both ICD-10 and DSM-IV are currently under review (Riemann et al., 2011). Contemporary psychiatry has been greatly influenced by these nosographic changes. In 2010, the DSM-V proposed criteria were tentatively manifested, which might be espected to increase the significance of the notion of sleep disturbances, after the presumable publication of DSM-V in 2013. It is under discussion whether the category primary/secondary insomnia should be replaced by the term « insomnia disorder ». In any case, this would emphasize the independence of the category in favor of the insomnia comorbidity concept, as suggested by the State of the Science conference on insomnia (NIH, 2005).

The concept and diagnostic criterion of insomnia are still fluctuating. In orde to become familiar with the current nosographic controversy, we show Table 1, which explains the general criteria for insomnia in ICSD-2, as well as the draft criteria for insomnia disorder in DSM-5 draft published in 2010 (Proposed DSM-5 Draft, 2010).

ICSD-2 General Criteria for insomnia (2005):

A. A complaint of difficulty initiating sleep, difficulty maintaining sleep, or waking up too early or sleep that is chronically unrestorative or poor in quality. In children, the sleep difficulty is often reported by the caretaker and may consist of observed bedtime resistance or inability to sleep independently.
B. The above sleep difficulty occurs despite adequate opportunity and circumstances for sleep.
C. At least one of the following forms of daytime impairment related to the nighttime sleep difficulty is reported by the patient: fatigue or malaise; attention, concentration, or memory impairment; social or vocational dysfunction or poor school performance; mood disturbance or irritability; daytime sleepiness; motivation, energy, or initiative reduction; proneness for errors or accidents at work or while driving; tension, headaches, or gastrointestinal symptoms in response to sleep loss; concerns or worries about sleep.

DSM-5 proposed Insomnia Disorder (2010).

A. The predominant complaint is dissatisfaction with sleep quantity or quality made by the patient (or by a caregiver or family in the case of children or elderly).
B. Report of one or more of the following symptoms:
   - Difficulty initiating sleep; in children this may be manifested as difficulty initiating sleep without caregiver intervention
   - Difficulty maintaining sleep characterized by frequent awakenings or problems returning to sleep after awakenings (in children this may be manifested as difficulty returning to sleep without caregiver intervention)
   - Early morning awakening with inability to return to sleep
   - Non restorative sleep
   - Prolonged resistance to going to bed and/or bedtime struggles (children)
C. The sleep complaint is accompanied by significant distress or impairment in daytime functioning as indicated by the report of at least one of the following: fatigue or low energy; daytime sleepiness; cognitive impairements (e.g., attention, concentration, memory); Mood disturbances (e.g., irritability, dysphoria); behavioral problems (e.g., hyperactivity, impulsivity, aggression); impaired occupational or academic function; impaired interpersonal/social function; negative impact on caregiver or family functioning (e.g., fatigue, sleepiness).
D. The sleep difficulty occurs at least three nights per week.
E. The sleep difficulty is present for at least three months.
F. The sleep difficulty occurs despite adequate age-appropriate circumstances and opportunity for sleep.

* Duration: i. Acute insomnia (< 1month); ii. Sub acute insomnia (1-3 months); iii. Persistant insomnia (> 3 months).
* Clinically Comorbid Conditions: i. Psychiatric disorder; ii. Medical disorder; iii. Another disorder.

Table 1. General criteria for insomnia in ICSD-2 (2005) and insomnia disorder in proposed DSM-5 (2010)

2.2 Japanese general population sample, re-analysed

Following this current insomnia concepts, we reanalyzed our Japanese population representative sample of 24,551 adults performed in 2000 (Abe et al., 2011). The present study was conducted using partial data from the Active Survey of Health andWelfare performed in June 2000 by the Ministry of Health, Labour and Welfare. To provide a representative sample of the general population in Japan, the survey was conducted through public health centers in 300 target areas randomly selected from the 881, 851 national census areas nationwide. The self-administered questionnaire consisted of 44 items covering the general health status, physical and psychological complaints and sleep habits and problems. We first selected cases reporting the presence of both insomnia symptoms and physical/psychological complaints during the past one month, identified based on the responses to the survey questionnaire about sleep problems and daytime functioning during the past one month. Then we excluded cases reporting a common comorbid sleep disorder (sleep-disordered breathing and restless leg syndrome).

The result was that we found a fairly high prevalence of insomnia (43.4%) as defined in this study (see Table 2) compared to before in the general population sample in Japan. Although previous studies have pointed out that Japanese people tend to underreport their sleep problems, because of cultural reticence compared with those in Western cultures, our results did not necessarily align with these studies (Abe & Mishima, 2008; Abe et al., 2011).

Possible reasons for the higher prevalence of insomnia obtained in our study include the following. First, following the ICSD-2 criteria, an item on "nonrestorative sleep" was added to our definition of insomnia. Secondly, our sample may have included cases with short-term insomnia occurring in less than the past one month (e.g. adjustment insomnia) in the absence of specifications on the duration and frequency of insomnia symptoms. The case definition of insomnia based partially on the ICSD-2 and DSM-IV was more liberal than the original definitions of the disorder. Lastly, the greatest factor responsible for such a higher prevalence rate was the inadequate assessment of daytime impairments associated with insomnia (Ohayon & Lemoine, 2004).

	Insomnia			Insomnia comorbid with depression		
Age class (years)	Subtotal % (*n*)	Male % (*n*)	Female % (*n*)	Subtotal % (*n*)	Male % (*n*)	Female % (*n*)
20–29	37.1 (1661)	33.4 (716)	40.7 (945)	5.6 (252)	4.9 (106)	6.3 (146)
30–39	41.7 (1881)	38.8 (834)	44.4 (1047)	4.4 (198)	3.3 (72)	5.3 (126)
40–49	41.5 (1911)	42.3 (953)	40.6 (958)	5.1 (236)	4.5 (102)	5.7 (134)
50–59	45.5 (2290)	45.1 (1107)	45.8 (1183)	5.0 (253)	4.6 (112)	5.5 (141)
60–69	48.1 (1653)	46.1 (780)	50.0 (873)	4.5 (155)	4.6 (78)	4.4 (77)
≤70	50.3 (1257)	48.8 (488)	51.3 (769)	10.5 (263)	9.0 (90)	9.8 (173)
Total	43.4 (10653)	41.7 (4878)	44.9 (5775)†	5.5 (1357)	4.8 (560)	6.2 (797)†

†Significant difference between men and women ($P < 0.001$, chi-square test).

Table 2. Presence of insomnia and insomnia comorbid with depression, by age, group and sex in a sample of the general Japanese adult population, conducted in 2000 (n=24, 551) .

It is possible that the complaints from participants were related to physical or psychological problems, which are separate issues from insomnia. However, as far as we know, there is no validated self-reporting tool about which researchers are in consensus for accurately measuring daytime impairments due to insomnia (Ohayon & Lemoine, 2004; Shekleton et al., 2010). One of the main reasons for this overdiagnosis of insomnia is that we used « daytime impairment » related to insomnia, including various items such as fatigue. This result implies one important subject, that is to say, "fatigue" itself can be regarded as a core symptom of insomnia (Choquet et al., 1993; Riemann et al., 2011). In recent literature, daytime sleepiness, hypersomnia and fatigue are common symptoms of depression (Franzen & Buysse, 2008). But, such symptoms can occur independently, or they may occur secondarily to insomnia comorbidity, as well as short- or long-term side effects of antidepressant medications themselves (Riemann et al., 2011).

3. The current psychobiological model of insomnia: Hyperarousal model

Although pathophysiology of insomnia remains to be explored, physiological hyperarousal evidenced by cognitive, endocrine, and neurophysiologic variables has been revealed to be involved in onset and development of insomnia (Bonnet, 1995, 2010; Riemann et al., 2010). Patients with insomnia suffer from cognitive deficit. Characteristically, they report their sleep and psychoperformance to be worse than are objectively measured (Endo, 1962; Orff et al., 2007). This "perceived" deficit is exactly what aggravates the QOL of insomniac patients and let them fall in a vicious cycle (Abe & Mishima, 2008). Also, insomnia is often induced by stressful events, and is assumed to develop by the 3P model (predisposing, precipitating and perpetuating factors), proposed by Spielman that is widely used to explain the onset mechanism of insomnia (Spielman et al., 1987; Ellis et al., 2011).

There is a need for clarification of pathophysiology of insomnia for development of efficient treatment skills and critical prevention of chronic insomnia. Just recently, reductions in hippocampal volume size have been reported in patients suffering from primary insomnia in brain research (Riemann et al., 2009). In the light of neurobiological theories of sleep-wake regulation, insomnia may be conceptualised as the final common pathway of the interaction of a genetic vulnerability to an imbalance between arousing and sleep-inducing brain centres, which is triggered by psychosocial and/or medical stressors, with perpetuating mechanisms such as maladaptive behaviours, learned sleep-preventing associations and cognitive factors (Basta et al., 2007; Riemann et al., 2010).

3.1 Development of chronic insomnia: 3P model

According to Spielman, insomnia is often induced by stressful events, and assumed to develop by his 3P model (predisposing, precipitating and perpetuating factors), that is widely used to explain the onset mechanism of insomnia (Spielman et al., 1987). Factors leading to the onset and worsening of insomnia are multidimensional in nature, and many life events and life stresses can result in acute insomnia. Inadequate stress coping behavior also precipitates insomnia, and heightens uneasiness and tension around being unable to sleep, thereby perpetuating the sleeplessness (Abe & Mishima, 2008). Furthermore, insomniacs may often engage in poor sleep hygiene, such as having an inadequate sleep environment, lack of daytime activities, and excessive afternoon napping. It is reported that the majority of people with insomnia attempt to cope with sleep problems in various ways,

have fewer adaptive coping skills, rely more on emotion-focused coping strategies than on problem-solving strategies and report lower feelings of mastery (Vollath et al., 1989). Reduced quality of life associated with insomnia has already been reported in a general population sample (LeBlanc et al., 2007).

3.2 Brief empirical evidence about insomnia

Since the classical study of Monroe *et al*., the validity of the hyperarousal concepts in patients with insomnia has been tested by measuring autonomous variables, including ECG-derived heart rate and heart rate variability, body temperature, whole-body metabolism and galvanic skin response (Riemann et al., 2010). The majority of studies measuring such variables in insomnia documented an increased arousal tone in this patient group. However, it is still unclear whether increased autonomic activity is causing insomnia or whether vice versa, insomnia and its sleep loss triggers increased autonomic activity.

Bastien *et al.* investigated a group of 285 patients evaluated for insomnia at a sleep medical clinic and found that 35% had a positive history for a sleep disturbances (Bastien & Morin, 2000). Dauvilliers *et al.* described that of 77 patients with primary insomnia, 72.7% reported familial insomnia compared to 24.1% in a non-insomnia control group in a French population sample (Dauvilliers et al., 2005). Similar result was reported by Morin's group from a Canadian polulation sample (Beaulieu-Bonneau et al., 2007). Drake *et al.* suggested that 37% of the variance in vulnerability to stress-related insomnia in siblings could be explained by familial aggregation (Drake et al., 2008).

Neuroimaging studies in insomnia are now widely used in human basic sleep research.. A PET study, condulted by Nofzinger *et al*., acquired data from 7 chronic insomniacs and 20 good sleeper controls during wakefulness and during consolidated Non-REM sleep. Patients with insomnia exhibited increased global glucose metabolism during wakefulness and Non-REM sleep. Patients with insomnia exhibited smaller declines in relative glucose metabolism from wakefulness to Non-REM sleep in wake promoting regions including the ascending reticular activation system. Reduced relative metabolism in the prefrontal cortex was found in insomniac while awake (Nofzinger et al., 2004). Another recent pilot study, using manual morphometry of structural magnetic resonance images showed that out of several regions of interest only one significant difference concerning a bilateral reduction of hippocampal volumes was found between 8 chronic insomniacs and 8 healthy control sleepers (Riemann et al., 2007). It remains to be determined whether these alterations of hippocampal structures are directly related to the insomnia. Nevertheless, these studies referred to above have taught us that the development of chronic insomnia is associated with measurable alterations of brain function pointing to Central Nervous System hyperarousal with a vulnerable familial aggregation.

The study of daytime performance in patients with insomnia has been driven by the assumption that short-term or chronic sleep loss has a negative impact on daytime functioning. Thus, such a compensatory effot might play an important role in the opposing effects of sleep deficits and hyperarousal that influence daytime performance.

The study of Orff *et al.* showed no impairments at all objective measures of cognitive performance in insomnia patients, with a discrepancy between subjective reports of deficits and objective neuropsychological tests (Orff et al., 2007). Despite the fact that there might be

only minor deficits in this population, investigating neuropsychological tests in large sample sizes might reveal stable deficits in the insomnia patients population (Edinger et al., 2008).

4. Symptomatic overlap: Insomnia-depression-anxiety connection

4.1 Insomnia-depression

In terms of descriptive symptomatology, insomnia symptoms often coexist with depressive and anxiety symptoms. As many as 90% of patients with depression will have sleep quality complaints (Tsuno et al., 2005). Alongside insomnia being the most common symptom of depression and anxiety disorder, persistent insomnia is a risk or exacerbating factor of depressive disorders (Riemann et al., 2003).

In a Japanese general population sample, the presence of insomnia comorbid with depression was 5.5% with a rate of 12.7% among the sample of people with insomnia (Abe et al., 2011). In line with this, Ford & Kamerow reported 14.0% as a prevalence of insomnia co-occurring with depression in a study based on 7954 American households (Ford & Kamerow, 1989). These studies showed that the frequency of insomnia comorbid with depression observed in Western countries is stable among Japanese adults as well (approximately one seventh of the insomnia population). Vollarath *et al*. state that insomnia constitutes an independent syndrome (Vollath et al., 1989), and Buysse *et al*. suggest that insomnia and depression are commonly comorbid, and insomnia comorbid with depression is an important intermediate phenotype (Buysse et al., 2008). Following the current insomnia-depression literature, we can consider as follows: (1) Insomnia and depression are bidirectionally related; (2) Insomnia is a risk factor for developing depression and (3) Insomnia is a risk factor for poor depression outcomes. Taken together, treating insomnia may favorably impacts the trajectory of depression (Franzen & Buysse, 2008).

4.2 Insomnia-anxiety

In general, insomniacs manifest their multi-complaints and they often have a comorbidity with anxiety disorders. Harvey *et al.* proposed her cognitive model of insomnia, explaining that excessive worriness about insomnia itself exacerbates insomnia. Bader *et al.* suggested that adverse childhood experiences are associated with sleep in primary insomnia (Bader et al, 2007), and Gregory *et al.* reported that familial conflicts in childhood predicted later insomnia, a modest but robust longitudinal link between family conflict during childhood and insomnia experienced at 18 years of age. (Gregory et al., 2006).

With regard to the comorbidity with anxiety disorder, the potential pathological link between insomnia and PTSD (posttraumatic stress disorder) and alcohol dependence should be more investigated. Especially, sleep disturbances have been considered the hallmark of PTSD for decades. Since insomnia has been observed in 90% of PTSD cases, both pharmacologic and psychosocial context of sleep of trauma should be needed to improve comorbid insomnia (Hendin et al., 2008).

4.3 Case Study 1; PTSD coexist with sleep disturbances

Clinically, sleep disturbances are common among individuals with posttraumatic stress disorder (PTSD), which are often resistant to first-line recommended treatments

(Singareddy & Balon, 2001). Recently, many studies and clinical experiences have suggested that sleep disturbances mainly representing insomnia and nightmare, have a distinct risk of suicide (Nadorff et al., 2011). If not, as some Holocaust survivors presented, impaired sleep and frequent nightmares had been considerable problems, even 45 years after the liberation (Rosen et al., 1991).

This PTSD patient, a 40 year-old careered woman, still suffered from her residual sleep disturbance, even if she partially recovered from her PTSD symptoms and improved her quality of life again, resulting in returning to her work environment. She was firstly presented an anxiety related with insomnia in the context of the accidental loss of her husband in front of her, at the age of 36. This event led her to consult a psychiatrist for the first time, and she has continued to be treated with medication and an individual psychotherapy regularly once a month. Her insomnia, nightmare and occasional suicidal ideation made her continue to maintain her treatment. Of importance, this patient exacerbated suicide ideation every year the day of incident approached. Outside her stabilized period, every time the clinician tried to reduce her nocturnal treament, she exacerbated her sleep complaints and related somatic complaints, alluding to the clinician her suicidal ideations.

For more than three decades, sleep disturbance had been considered the hallmark of PTSD (Hendin et al., 2008; Nadorff et al., 2011). Since insomnia has been observed in 90% of PTSD cases and nightmare related to the trauma in 70%, this is understandable (Hendin et al., 2008). In this case, the clinician has mainly prescribed paroxetine (10-20 mg) and trazodone (25-50 mg) at night to improve subjective sleep disturbances. One of the paradoxical difficulties in psychopharmacology is that there has been increasing awareness of psychotropic-related sleep disruptions in PTSD patients. Especially, it is reported that selective serotonin reuptake inhibitors (SSRIs), usually prescribed as a first-line medication to PTSD, have conversely been associated with clusters of side effects, including insomnia and nightmare symptoms (Li et al., 2010). Trazodone, prescribed at low dose, may reverse the SSRI-induced insomnia; increases the antidepressant effects of SSRIs; promotes sleep through its sedative properties; and suppresses rapid eye movement sleep, thus reducing nightmares associated with PTSD (Sungareddy & Balon, 2001). This case showed that the residual symptoms related with sleep in PTSD resisted, even though the traumatic event had passed away and the patient recovered on a social function level. Probably, most experienced psychiatrists must have had the same treatment impressions before. It is true that sleep disturbance should be more than a marker of PTSD and hence may be important in the identification of suicidal ideation (Nadorff et al., 2011). Recently, Hendin et al. (2008) have insisted on the equal importance of the psychosocial context of trauma in treating sleep disturbance associated with PTSD. It is stressed again that sleep assessment should be considered in the evaluation of suicide risk in PTSD. Both pharmacological and psychotherapeutic approaches to the disorder have concentrated on improving sleep complaints. This case showed us the necessity of long-term sleep-focused approach in order to treat patients suffering from PTSD with suicidal ideation. That implies that incorporating individual psychotherapy, combined with sleep hygiene approach, can lead the patient to recovery from traumatic event in the long term setting. The emotional consequence of suicide will be devastating to the victim's family, friends, community, and society. Studies of incidence, risk and protective factors related to sleep disturbances need to be high on the research agenda across many countries.

4.4 PSQI-A scale

As this case description shows, PTSD patients report a wide variety of subjective complaints. These subjective sleep disturbances are non-specific and also observed in other sleep disorders and psychiatric clinical samples. For example, PTSD and depressed patients show similar global score on The Pittsburgh Sleep Quality Index, one of the most frequently used self-report instruments to assess sleep quality. Disruptive nocturnal behaviors (DNB), such as trauma-related nightmares, may represent more specific sleep disturbances in PTSD. Recently, Germain *et al.* developed the PSQI Addendum for PTSD (PSQI-A), a brief sleep scale for PTSD, to evaluate DNB (Germain et al., 2005). This self-report instrument consists of 7 items that focus on the frequency of seven DNB, and includes three additional items regarding the frequency of anxiety and anger accompanying DNB and the timing of these events during tht night (Table 3). Such an assessment may support the clinical utility of assessing DNB to determine the need for further PTSD evaluation and intervention.

PSQI Addendum for PTSD
1. During the past month, how often have you had troubles sleeping because you…
a. Feel hot flashes:
b. Feel general nervousness:
c. Had memories or nightmares of a traumatic experience:
d. Had severe anxiety or panic, not related to traumatic memories:
e. Had bad dreams, not related to traumatic memories:
f. Had episodes of terror or screaming during sleep without fully awaking:
g. Had episodes of "acting out" your dreams, such as kicking, punching, running , or screaming:
2. If you had memories or nightmares of a traumatic experiences during sleep
a. How much anxiety did you feel during the memories/nightmares?
b. How much anger did you feel during the memories/ nightmares?
c. What time of night did most memories/nightmares occur?

Table 3. Pittsburgh Sleep Quality Index Addendum for PTSD (PSQI-A) (Germain et al., 2005)

4.5 Stress coping and sleep hygiene among Japanese people with insomnia

Factors leading to the onset and worsening of insomnia are multidimensional in nature, and many life events and life stresses can result in acute insomnia. Inadequate stress coping behavior also precipitates insomnia, and heightens uneasiness and tension around being unable to sleep, thereby perpetuating the sleeplessness. Furthermore, insomniacs may often engage in poor sleep hygiene, such as having an inadequate sleep environment, lack of daytime activities, and excessive afternoon napping (Abe et al., 2011). The majority of people with insomnia attempt to cope with sleep problems in various ways, have fewer adaptive coping skills, rely more on emotion-focused coping strategies than on problem-solving strategies and report lower feelings of mastery (LeBlanc et al., 2007). Reduced quality of life associated with insomnia has already been reported in a general population sample. We have recently studied specific daily stress coping behaviors (SCBs) and sleep hygiene practices (SHPs) of people with insomnia in our Japanese population based sample.

As a result, we clarified that Japanese adults with insomnia might also engage in various maladaptive SCBs and SHPs (Table 4). Most importantly, we found that people with insomnia may not necessarily engage in the same SCB as insomniacs comorbid with depression (Abe & Mishima, 2008). It has often been considered that treatment with insomnia played a bunch of treatments of depression. But, our findings indicated that novel therapeutic strategies need to be developed, taking into account both characteristics of insomnia and depression. These kinds of concrete findings about daily behaviors related with insomnia may offer critical insights for developing effective sleep educational preventive programs in public health, as reported by Morin's group in Canada (Morin et al, 2006). For example, concerning substance dependence, the association between insomnia and its self-medication with alcoholism has been acknowledged (Brower et al., 2001). Our unpublished data in alcoholic groups in Japan also showed that the majority of middle-aged alcoholic patients entering treatment reported insomnia symptoms and recognized themselves their diminished quality of sleep (Asami et al., 2011).

	Insomnia (n = 10653)					Insomnia comorbid with depression (n = 1357)				
		Crude		Adjusted†			Crude		Adjusted†	
	N	OR	95%CI	OR	95%CI	N	OR	95%CI	OR	95%CI
Stress coping behaviors (SCB)										
Bearing the stress without taking any action (Bearing)	1576	1.97	1.78–2.18	1.69	1.52–1.88	378	3.49	2.96–4.10	3.44	2.92–4.05
Smoking (Smoking)	1954	1.22	1.12–1.33	1.26	1.15–1.38	317	1.48	1.24–1.76	1.73	1.44–2.08
Eating something (Eating)	1663	1.27	1.16–1.39	1.22	1.11–1.34	273	1.58	1.33–1.88	1.51	1.26–1.81
Watching TV/Listening to radio (TV/ Radio)	3650	1.26	1.17–1.35	1.18	1.10–1.27	537	1.57	1.35–1.83	1.52	1.30–1.78
Making an effort to solve problems actively (Problem solving)	1609	0.88	0.80–0.96	0.87	0.80–0.95	121	0.50	0.39–0.64	0.50	0.39–0.65
Taking it easy (Ease)	3630	*n.s.*	–	*n.s.*	–	354	0.72	0.61–0.85	0.74	0.63–0.87
Making plans to take time off (Time off)	734	*n.s.*	–	*n.s.*	–	65	*n.s.*	–	*n.s.*	–
Sleep hygiene practices (SHP)										
Drinking alcohol (Alcohol)	2961	1.24	1.15–1.34	1.27	1.18–1.38	349	*n.s.*	–	*n.s.*	–
Reading books/Listening to music (Books/ Music)	3747	1.20	1.12–1.29	1.24	1.15–1.33	460	1.36	1.16–1.59	1.39	1.19–1.63
Taking a bath (Bath)	4983	1.13	1.05–1.21	1.09	1.01–1.17	587	*n.s.*	–	*n.s.*	–
Trying to have regular daily habits (Regularity)	4114	*n.s.*	–	*n.s.*	–	420	0.69	0.59–0.80	0.64	0.55–0.75
Taking light exercise (Exercise)	2174	*n.s.*	–	*n.s.*	–	239	*n.s.*	–	*n.s.*	–

†Adjusted for sex, age, and presence of stress by multiple logistic regression analyses. CI, confidence interval; Crude, non-adjustment; OR, odds ratio ($P < 0.01$).

Table 4. Stress coping behavoir and sleep hygien practices in the Japanese general adult sample

Abe *et al*. have recently studied several specific daily stress coping behaviors and sleep hygiene practices of people with adult insomnia in the Japanese adult general population (Abe et al., 2011). As a result, they clarified that Japanese adults with insomnia might also engage in various maladaptive conducts. They also found that people with insomnia may not necessarily engage in the same behaviors and practices as insomniacs comorbid with depression. Although this study mainly targeted adults, future research needs to examine these aspects among minors in order to clarify the onset of insomnia and its temporal development into chronic adult insomnia. Such minors may be characterized by vulnerabilities in how they perceive and experience stressful life events negatively during adolescent periods. Most of them are not seeking help, thus possibly they will continue to engage in self-help maladaptive practices, such as substance abuse, until they are finally diagnosed with chronic insomnia or depression later (Vollath et al., 1989; Wong et al., 2009).

4.5.1 Stress coping behaviors among people with insomnia

As far as we know, our study is the first report that investigates stress-coping behaviors among people with insomnia in the general adult population. According to the classical formulation by Lazarus and Folkman (1984), coping behavior refers to cognitive and behavioral efforts to manage external and internal demands (Morin & Espie, 2004). There are two types of coping behaviors: problem-focused and emotion-focused behaviors. With regards to the coping behaviors among people with insomnia, Morin *et al*. indicate that, compared with good sleepers, people with insomnia are apt to perceive their lifestyle as more stressful and choose more emotion-focused coping behaviors (Morin et al., 2003). This does not contradict reports indicating that people with insomnia tend to internalize stress, affecting emotions (Basta et al., 2007). Similar trends were observed in the sample of people with insomnia in this study (Abe et al., 2011). Our multivariable logistic regression analysis revealed that, among the seven SCBs, insomnia was positively related to the emotion-focused coping behaviors of bearing, smoking, eating, and TV/radio. Bearing had the strongest positive correlation with insomnia (OR = 1.69), and an even stronger correlation with insomnia comorbid with depression (OR = 3.44). Therefore, this study indicates that problem-focused behaviors represented by Problem-solving could be helpful in overcoming insomnia. While Ease was not significantly related to insomnia, it had a significant relation with insomnia comorbid with depression (OR = 0.74). This indicates that people with insomnia may not necessarily engage in the same stress-coping behavior as insomniacs comorbid with depression. The present findings indicate that novel therapeutic strategies need to be developed, taking into account both characteristics of insomnia and depression. This study further revealed a strong positive association between Smoking and insomnia (OR = 1.26). Previous research in Europe and in the United States indicates a relationship between nicotine consumption through smoking and poor sleep quality (Morin & Espie, 2004). Furthermore, the strong association between Smoking and insomnia comorbid with depression (OR = 1.73) indicates that individuals with insomnia comorbid with depression tend to rely on more unhealthy coping strategies in their daily life. Our results might highlight the importance of strongly urging people complaining of insomnia to quit smoking. Eating was significantly related to insomnia. A previous epidemiological study reported that irregular eating habits and subjective sleep insufficiency were closely associated. TV/Radio is also significantly related to insomnia. Morin *et al*. indicated that many individuals initiate a variety of self-help strategies to alleviate insomnia, including listening to music and relaxation (Morin et al, 2006). In fact, these individuals may

experiment with a variety of these passive emotional focused self-help remedies for a considerable period of time before seeking professional help (Morin & Espie, 2004).

4.5.2 Sleep hygine pracitices among people with insomnia

There have been several studies that have shown that individuals with insomnia often engage in some inappropriate sleep practices. In a population-based sample of 258 insomniacs, Jefferson *et al*. reported that, compared with healthy people, insomniacs more habitually drank alcohol before going to bed (Jefferson et al., 2005). Our study also demonstrated that alcohol consumption before going to bed is positively related to insomnia. Research in the United States suggests that drinking alcohol is an important risk factor for sleep problems. In their comparison of sleep habits among people in ten different countries, Soldatos *et al*. found that Japan ranked the highest in terms of the prevalence of alcohol use as a sleep aid (30.3%) (Soldatos et al., 2005). Thus, it is critical to provide sleep hygiene education about minimizing alcohol consumption before bedtime to people with insomnia. Our analysis further found that Books/Music was also positively related to insomnia. Some previous studies have reported that reading behavior is significantly more frequent among groups with insomnia than control groups. Morin *et al*. found in their epidemiological survey of a general population in Canada that insomnia syndrome sufferers use music (OR = 2.6) and reading (OR = 1.8) as self-help strategies to facilitate sleeping (Morin et al. 2006). In our study, combining Books and Music into one item in the questionnaire may have comparatively reduced the odds ratio (Table 4). One epidemiological study among Japanese indicates that poor exercise habits are associated with insomnia. Based on this finding, we hypothesized that physical activity would be an inhibiting factor for insomnia symptoms; however, there was no significant relationship between Exercise and insomnia. Previous research suggests that daytime physical activity improves sleep. The inconsistency in the findings might be attributable to the lack of information available regarding the type (level), duration, and frequency of physical activity in our study. While Bath was slightly related to insomnia, it had no significant association with insomnia comorbid with depression. Subjective sleep sufficiency is better for individuals when they take a bath before going to bed rather than when they do not. Taken together, these observations may indicate that taking a bath improves the subjective quality of comorbid depression. By contrast with previous studies, our analysis found no significant association between Regularity and insomnia. This may be attributable to the fact that we did not define the behaviors belonging to this SHP in a concrete manner. Regular exposure to photic and nonphotic time cues (Zeitgebers) for the circadian clock system supposedly stabilizes the acrophases of the sleep–wake rhythm as well as the physiological rhythm, allowing one to fall asleep and maintain sleep more easily (Wirz-Justice et al., 2009). The strong negative association between Regularity and insomnia comorbid with depression (OR = 0.64) found in the present study supports a treatment emphasis on regularity for mood disorders including bipolar disorder.

4.6 Future remarks about insomnia

The studies of the relationship between insomnia and suicidality started from investigating the relationship between depression and suicidality. It is still needed to clarify whether insomnia could be a distinct factor related to suicidality, even controlling for depression (Pigeon & Caine, 2010). Suicide prevention of depression often includes insomnia, but they are not always in line and insomnia has a distinct psychopathology, different from the one of depression. According to current etiological models of insomnia, a cognitive, emotional

and physiological hyperarousal may play an important role in the development and maintenance of the disorder. This hyperarousal concept has just recently been summarized in several review articles. Riemann *et al.* pointed out that it is important to note that two effects are, at least to some extent, opposing in chronic insomnia: on the one hand, sleep deficits or chronic minor sleep loss affects neurobiological processes and neuropsychological performance; on the other hand, there is the elevated arousal level which can be measured in several physiological systems (Riemann et al., 2011). These opposing processes might construct the rather paradoxal psychopathology of insomnia (Baglioni et al., 2010; Riemann et al., 2010, 2011). Needless to say, further studies will be needed to clarify the relationship between insomnia related hyperarousal and suicidality. Perhaps, subtyping insomnia patients according to signs of hyperarousal and the intensity of daytime impairment, such as the intensity of fatigue, might offer a way to disentangle the pathology of suicidality. In this sense, attachment theory may provide a useful framework for considering how the socio-emotional climate influences affect and arousal across the lifespan, and may be particularly important for understanding psychopathology of insomnia. For example, anxious attachment styles which are characterized by 'hyper-activating' strategies during times of threat or stress, may predispose an individual to insomnia by influencing stress-arousal systems and cognitions related to the emotional and physical availability of the partner (Troxel & Germain, 2011).

5. Adolescent insomnia

Adolescents experience changes in their opposing societal demands, such as early school-start times and an increase in the significance of social roles coincide with these physiologic changes (Brand & Kirov, 2011; Liu & Buysse, 2006). These incongruous demands may explain why adolescents are prone to sleep disturbances, such as delayed phase sleep syndrome and insomnia. The multiple changes that adolescents experience can be very stressful, and serve as precipitating factors that activate biological and/or psychological diathesis, and subsequently, to the development of other mental health problems.

Studies in adults have already found that insomnia is associated with psychological problems (Singareddy & Balon, 2001). However, little research has explored the relationship between insomnia and mental health during adolescence and young adulthood. Substantially less research has evaluated insomnia and psychological disorders in adolescents. Safer D.J. suggested that adolescents differed from adults in suicidal behavior in their greater attempt rate, higher attempt/completion ratio, and lower rates of short and intermediate completion following psychiatric treatment. He claimed that the frequent practice of combining adult and adolescent suicide and suicide behavior findings can result in misleading conclusions (Safer, 1997).

5.1 Epidemiology: Prevalence of adolescent insomnia

Youth and adolescent suicidality constitutes a major public health problem, ranking among the leading causes of death for young people in many countries worldwide. Risk for completed suicide increases dramatically during adolescence, and research implicates an array of associated factors from genetic, biological, psychosocial, and cognitive domains (Bridge et al., 2006; Brand & Kirov, 2011). Sleep disturbances are prevalent not only among adults but also among 10–40% of adolescents (Liu et al., 2000; Johnson et al., 2006; Roane & Taylor, 2008). An estimated 10.7% of adolescents in the general population experience insomnia according to

DSM-IV criteria (Johnson et al., 2006). Roane & Tayler also showed that insomnia symptoms were reported by 9.4% of the 4495 adolescents, 12 to 18 years old, suggesting that one out of ten adolescents met the criteria for insomnia (Roane & Taylor, 2008). The authors examined adolescent insomnia as a risk factor for mental health problems in a longitudinal study. They concluded that insomnia should be treated with specific interventions as an independent disorder in adolescents (Taylor & Roane, 2010). The Japanese research team of Ohida *et al*. has performed large-scale epidemiological studies on the sleep status of Japanese adolescents (Ohida et al., 2004; Kaneita et al., 2006). In a survey of approximately 106,300 Japanese junior and high school students, 30.6% reported an average sleep duration of less than 6h per night. Of these, 12.5% reported excessive daytime sleepiness, and 40% were not satisfied with their sleep quality (Ohida et al., 2004). Another survey reported that 23.5% of adolescents experienced symptoms of insomnia (Kaneita et al., 2006). Most studies of sleep disturbances among adolescents have focused on sleep deprivation and insomnia, and other types of sleep disturbances have not been adequately addressed.

5.2 Adolescent insomnia and suicidality

Sleep undergoes substantial changes during adolescence and suicide risk begins to increase during this period as well (Liu & Buysse, 2006; Wong et al., 2011). Adolescent sleep is characterized by widespread sleep restriction, irregular sleep schedules, daytime sleepiness, and elevated risk for sleep disturbances (Gangwisch et al., 2010). Sleep is indispensable in terms of brain maturation and learning for adolescents. Maladaptive sleep habits prevent them from growing, even run a risk of increasing suicide ideation. Sleep loss or disturbances are likely to signal an increased risk of future suicidal action in adolescents. Large-scale prospective studies and neurobiological studies are needed for a better understanding of the complex relationship between sleep, psychopathology, and youth suicidal behavior.

Research with adolescents has demonstrated a clear relationship between suicidal ideation and sleep problems. Cross-sectional studies have found that adolescents with insomnia experience more depressive symptoms, and suicide ideations and attempts and are more likely to use alcohol, cigarettes, illicit drugs, or a combination of these substances. In a provident epidemiological study of French teenagers, Choquet *et al*. found that adolescents with suicidal ideation reported more insomnia as well as more nightmares than adolescents who denied suicidal ideation (Choquet & Menke, 1990). In their subsequent study, suicidal ideation was linked to more sleep difficulties and frequent feelings of daytime tiredness (Choquet et al., 1993). It follows that the findings linking sleep disturbance with suicidality may serve as a proxy for severity of insomnia comorbid with depression more generally.

Better understanding the relationship between disturbed sleep and suicidality in adolescents may also serve for suicide prevention with this population (Goldstein et al., 2008). There is evidence to suggest that some factors associated with adolescent suicide may be different from adult suicide (Safer, 1997). For example, although impulsive–aggressive behavior is a common risk factor for both adult and teenage suicide, aggression and impulsivity are traits highly related to suicidal behavior in adolescents (Apter et al., 1995). Higher levels of impulsive aggressiveness play a greater role in suicide among younger individuals with importance decreasing with age. Adolescents with aggression and conduct disorders may be suicidal even in the absence of depression. Psychosocial factors associated

with adolescent suicide, such as stress and contagion, bullying and peer victimization may also be different from adults. Alcohol and drug abuse contribute significantly to the risk of suicide in teenagers (Apter et al., 1995). Additional potential contributors to suicidal behavior in depressed adolescents are early defined traits such as temperament and emotional regulation. One recent study suggests that suicidal youth are characterized by highly maladaptive regulatory responses and low adaptive emotional regulation responses to dysphoria (Tamas et al., 2007).

5.3 Sleep problems in highschoolers, students and youth sample

College students will be an ideal population to examine sleep disturbances and mental health relationships (Yang et al., 2003; Taylor et al., 2010; Nardoff et al., 2011). Yang *et al*. investigated the 1,922 first year college students' coping strategies for sleep disturbances and their effectiveness in Taiwan (Yang et al., 2003). They pointed out the relative lack of effective coping strategies for the management of such problems in this population. The results showed that taking naps and adjusting sleep schedules were coping strategies associated with better sleep quality. As mentioned throughout, the young adult age group is particularly susceptible to the onset of major psychiatric disorders. If so, the next logical step would be to develop primary and secondary sleep prevention programs for behavioral changes in this population (Liu & Buysse, 2006).

Brand *et al*. evaluated the effect of early stage intense romantic love on sleep quality in 113 adolescents (mean age: 17.8) (Brand et al., 2007, 2010). The research showed that adolescents reported significantly less daily sleepiness, higher daily concentration, more physical activity, and better mood compared to the other groups. Intense love in adolescents seems to be comparable with hypomanic state of bipolar mood spectrum. Intense positive emotions could disturb sleep quantity through the presence of heightened psychophysiological arousal, while improving perceived sleep quality and daytime activity. At least, combined PSG or actigraphic studies may be needed to understand the effects of such intense and positive emotions on sleep among adolescents and youth adults.

6. Another symptomatic aspect: Insomnia and nightmare, distinct suicide risk?

Clinical observations have showed that nocturnal sleep disturbances, including insomnia and recurrent nightmares, represent common distressing sleep complaints that might have important prognostic and therapeutic implications in psychiatric patients. Epidemiological studies have demonstrated that insomnia, nightmares, and sleep insufficiency are associated with elevated risk for suicide. Several studies have suggested an independent predictive role of nightmares in future suicidal behavior. It should be more noticed that nightmares may be more than a marker of PTSD and really important in the identification of suicidal ideation in primary care.

6.1 Sleep disturbances in mental health epidemiology

There is a consensus that one growing area of research in mental health includes the study of the relationship between sleep disturbances and suicidality in this decade (Ağargün & Beşiroğlu, 2005; Bernert et al., 2005; Bernert & Joiner, 2007; Pigeon & Caine, 2010). Increasing evidence in both clinical and epidemiological studies suggests that disturbances in sleep are

associated with an elevated risk for suicidal behaviors. Both sleep disorders and general sleep complaints appear to be linked to greater levels of suicidal ideation and depression, as well as both attempted and completed suicide (Fawcett et al., 1990; Ağargün et al., 1997; Krakow et al., 2000). As these provident studies have already stressed, one major expected suggestion is that sleep disturbances may have prognostic significance in predicting suicide among patients with depression. A recent study conducted in Japan, Fujino *et al.* showed that, among 13,259 middle-aged adults, only difficulty maintaining sleep (sleep maintenance insomnia), compared to other sleep disturbances (e.g., difficulty initiating sleep, nonrestorative sleep), significantly predicted death by suicide 14 years later (Fujino et al., 2005). But, depression was not accounted for when examining the association between sleep and completed suicide. Such findings would often elucidate whether sleep disturbances stand alone as a risk factor for completed suicide or, conversely, whether such sleep complaints simply vary with increased depressive symptoms (Ağargün & Beşiroğlu, 2005; Fujino et al., 2005). Sleep problems and more specifically, significant changes in sleep, have been considered as warning signs of suicide in many mental health policies. Thus, improvement in the identification of risk factors for suicidal behaviors and possible early intervention and postvention thus ultimately enhance our competence to intervene and prevent death by suicide (Krakow et al., 2011).

Fawcett *et al.* conducted the first study to prospectively examine sleep, depression and suicide in 1990 (Fawcett et al., 1990). They considered insomnia to be one of the 'modifiable risks' for suicide in patients with depression. Ağargün *et al.* demonstrated a significant association between poor sleep quality and suicidal behavior in depression (Ağargün et al., 1997). Further studies will be needed to the posssible intervention with regard to suicidality.

Again, does insomnia (sleep disturbances) still manifest distinct suicide risk, even controlling after several confounding factors? During several years, many studies and clinical experiences have tried to investigate this concern (Wojnar et al., 2009; Li et al., 2010; Pigeon & Caine, 2010). But, this question was already asked nearly one century ago by a British doctor. In 1914, in the medical journal Lancet, Pronger wrote an epoch-making article, entitled "Insomnia and Suicide" (Pronger, 1914). His clinical intuition still impresses us enormously, even about one century afterwards. A recent clinical case report stressed again that sleep assessment should be considered in the evaluation of suicide risk in depressed patients (Mahgoub, 2009).

6.2 Chronobiogical factors and diurnal fluctuation of suicidality

The study of chronobiological factors in the relationship between sleep and suicidal behaviors remains a largely unexplored, yet fruitful area of research (Ağargün & Beşiroğlu, 2005; Bernert & Joiner, 2007). A diurnal variation in the tiling of self-injurious behaviors and completed suicide is supported by several reports. Blenkiron *et al*. prospectively assessed 158 patients presenting at a hospital referred for psychiatric assessment due to deliberate self-harm (Blenkiron et al., 2000). The authors classified these deliberate self-harm incidents as suicide attempters, and concluded that the frequency of these acts were higher in the evening and lower in the early morning hours. They also showed a bimodal peak in frequency for deliberate self-harm among older and younger adults. And they concluded that the severity of deliberate self-harm appeared to vary according to the time of day (Blenkiron et al., 2000). In another study in Japan, ambulance report records were

retrospectively reviewed for a 7-year period in Tokyo to examine time-of-day and documented suicide attempts. Results indicated that suicide attempts showed a peak earlier in the evening (18h00) compared to the morning (Motohashi, 1990).

Selvi *et al.* assessed 80 patients clinically diagnosed with major depression and 80 healthy subjects who were demographically matched with the patient group (Selvi et al., 2010). Results showed that morningness-type circadian rhythm may play as a significant relief factor after the onset of major depression, but sleep variables of chronotype and sleep quality did not significantly predict suicide ideation after controlling for depressive symptoms in the major depression group. They concluded that suicide ideation and poor sleep quality were antecedents of depression symptom severity in patients with major depression. They discussed these findings under the theoretical assumptions concerning possible relations between chronotype, sleep quality, depression, and suicidality (Selvi et al., 2010).

In studying time-related risk factors, additional research is needed, particularly studies that better define the severity of suicidal behaviors. It will be important for such studies to carefully distinguish suicide attempts and deliberate self-harm with an intent to die from self harm behaviors without suicidal intent. Investigation of the timing of sleep and suicidal acts may inform risk assessment procedures, emergency responding and surveillance, as well as treatment (Bernert & Joiner, 2007). There is an association between circadian rhythms and suicidality. This topic has always been investigated in terms of diurnal fluctuation of symptoms related to depression or Seasonal Affective Disorder (Wirz-Justice et al., 2009). Future research will also be necessary to thoroughly evaluate chronobiological correlates of suicidality in non-clinical samples for preventative purposes.

6.3 Sleep homeostasis hypothesis and suicidality

Sleep abnormalities are common in patients with suicidal behavior. Sleep complaints such as insomnia, hypersomnia, nightmare, and sleep panic attacks are frequent in suicidal adolescents and adults. Results from school-based survey in the USA indicate that whereas insomnia and hypersomnia independently increase risk for suicidal ideation in adolescents, the presence of both insomnia and hypersomnia incurs further increased suicidal risk in this population (Roberts et al., 2001). In another study, a significant and temporal relationship between sleep problems and completed suicide has been observed (Goldstein et al., 2008). Considerable evidence supports a strong link between sleep disturbances and suicidality but the pathway remains to be established (Sher, 2008).

In 2003, an innovative theoretical model, called "sleep synaptic hypothesis", reflecting on the significance of slow-wave activity and its homeostatic regulation was proposed (Tononi & Cirelli, 2003). According to this hypothesis, neuroplastic processes occurring during wakefulness result in a net increase in synaptic strength in many brain circuits. The role of sleep is to downscale synaptic strength to a baseline level that is energetically sustainable, makes efficient use of gray matter space, and is beneficial for learning and memory. Thus, sleep is the price we have to pay for plasticity, and its goal is the homeostatic regulation of the total synaptic weight impinging on neuron (Tononi & Cirelli, 2003, 2006). It has been suggested that wakefulness is associated with synaptic potentiation in several cortical circuits; synaptic potentiation is tied to the homeostatic regulation of slow-wave activity;

slow-wave activity is associated with synaptic downscaling; and active synaptic downscaling occurring during sleep is beneficial for cellular functions and is tied to overnight performance improvement.

Hence, many aspects of behavioral performance improve after sleep and are negatively affected by sleep deprivation, and it is conceivable that avoiding synaptic overload by maintaining synaptic homeostasis would be beneficial for many cellular processes, such as energy metabolism and membrane maintenance. Clinically, sleep deprivation may affect fatigue complaints and the production of dreams, which is particularly important for adolescent's development. It is possible that disruption of synaptic homeostasis underlies sleep abnormalities, leading or contributing to suicidal behavior. Serotonergic mechanisms may affect sleep regulation, are implicated in the pathophysiology of suicidal behavior, and may be involved in the relation between sleep abnormalities, synaptic homeostasis and suicidal behavior (Sher, 2008). Taken together, theoretically, sleep difficulties should be considered in prevention and intervention effort for patients at risk for suicide. Prevention effort should target good sleep hygiene and early detection and treatment of problematic sleep patterns in order to decrease risk for suicide (Liu & Buysse, 2006). Better understanding of the relationship between disturbed sleep and suicidality may serve to inform effort for suicide prevention.

7. Nightmares

7.1 Nightmare; Definition and epidemiology

Both insomnia and nightmare showed classical and, at the same time, a novel symptomatic aspect in psychiatric epidemiology. Clinical observations have showed that nocturnal sleep disturbances, including insomnia and recurrent nightmares, represent common distressing sleep complaints that might have important prognostic and therapeutic implications in psychiatric patients. Epidemiological studies have demonstrated that insomnia, nightmares, and sleep insufficiency are associated with elevated risk for suicide (Hasler & Germain, 2009).

Dreams are a remarkable experiment in psychology and neuroscience, conducted every night in every sleeping person. They show that the human brain, disconnected from the environment, can generate an entire world of conscious experiences by itself (Nir & Tononi, 2010). Both DSM-IV and ICSD-2 criteria converge on defining nightmares as intensely disturbing dreams that awaken the dreamer to a fully conscious state and generally occur in the latter half of the sleep period (Table 5.).

Lifetime prevalence of nightmares in the general population is unknown, but large epidemiological studies indicate that about 85% of adults have experienced at least one nightmare within the past year (Levin & Nielsen, 2007). Further investigations suggest that the prevalence may almost approach 100%! The estimated frequency of clinically significant nightmares (occurring at least weekly) is 4–10% in the general population (Nielsen et al., 2006). Similar rates are reported from different cultures. There is a significant gender difference in nightmare frequency, with women of all ages reporting nightmares more frequently than men. Age is also relevant: nightmares are less frequent among the elderly (Levin & Nielsen, 2007; Nielsen et al., 2006).

Nightmare Disorder
A. Repeated occurrences of extended, extremely dysphoric and well-remembered dreams that usually involve efforts to avoid threats to survival, security or physical integrity and that generally occur during the second half of the major sleep episode.
B. On awakening from the dysphoric dreams, the person rapidly becomes oriented and alert.
C. The dream experience, or the sleep disturbance produced by awakening from it, causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.
D. The dysphoric dreams do not occur exclusively during the course of another mental disorder (e.g., a delirium, Posttraumatic Stress Disorder) and are not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition.

Table 5. DSM-5 proposed criteria for nightmare disorder ([84], 2010).

7.2 Nightmare; Etiology

Because nightmares are often, but not necessarily, associated with PTSD, many specialists distinguish post-traumatic and non-traumatic (idiopathic) nightmares (Hasler & Germain, 2009; Levin & Nielsen, 2009). Post-traumatic nightmares reflect the long-lasting effect of a wakeful traumatic experience, whereas the cause of non-traumatic nightmares is unknown. Numerous studies have found that nightmare frequency is associated with psychopathological symptoms (Levin & Nielsen, 2007), but because most of these studies do not strictly distinguish between post-traumatic and non-traumatic nightmares, the interpretation of the results is ambiguous. Levin & Nielsen described six broad psychopathological categories that are associated with nightmares: anxiety symptoms, neuroticism and global symptom reporting, schizophrenia-spectrum disorders, other psychiatric disorders, behavioral health problems and sleep disturbances, and PTSD (Levin & Nielsen, 2007). A common feature of these pathologies is notable waking emotional distress, suggesting that nightmares may play a role in processing of these experiences. The studied reviewed above also suggest that the connection between early experiences, brain development, and nightmare experiences might involve failures in emotion regulation (Nielsen et al, 2006; Levin & Nielsen, 2009). Current models of nightmare production seem to emphasize negative emotionality as having a central role in determining dream affects. Ağargün *et al.* previously reported that the prevalence of childhood traumatic experiences was higher among adult who ''often'' had nightmares than among adults who ''sometimes'' or ''never'' had nightmares (Ağargün et al., 2003). With regard to the associations between nightmares and mental health status, Nielsen *et al.* studied adolescents (aged 13–16) and reported a significant association between the frequency of nightmares and the level of anxiety (Nielsen et al., 2006). To date, very few studies have investigated the prevalence of nightmares in adolescents, compared to adults. In Japan, analyzing 90,081 nationwide adolescent sampled data, Munezawa *et al.* showed that the prevalence of nightmares was 35.2% among Japanese adolescents (more than one third) (Munezawa et al., 2011). The results of this study should be considered in the prevention of nightmares among Japanese

adolescents. They concluded that it is important to maintain regular sleep habits for preventing this symptom, and proposed that health education about regular sleep habits should be promoted among Japanese adolescents in a near future.

7.3 Nightmare and attachment

Interestingly, from the viewpoint of both attachment theory and epidemiology, Csóka *et al.* have hypothesized that adults who experienced early maternal separation (before one year of age and lasting at least one month) would report more frequent nightmares and bad dreams (Csóka et al., 2011). In the frame of the Hungarostudy Epidemiological Panel, 5020 subjects interviewed, significant associations were found between early maternal separation and both frequent nightmare experience in adulthood and increased frequency of oppressive and bad dreams. Current depression scores fully mediated the association between early separation and nightmares, but not the association between early separation and negative dream affects. The authors interpreted these findings as a trait-like enhancement of negative emotionality in adults who experienced early maternal separation. This enhancement influences the content of dreams and, when it takes the form of depression, also influences the frequency of nightmares. The effect of early maternal separation on nightmares and bad dreams is relevant, which merits further attention (Csóka et al., 2011).

7.4 Nightmare and suicidality

As we have mentioned above, frequent nightmares have been noted to be related to suicidality in depressed patients, particularly among women (Ağargün et al., 1998). A prospective follow-up study in a sample drawn from the general population also reported that the frequency of nightmares is directly related to the risk of suicide (Tanskanen et al., 2001; Turvey et al., 2002; Bernert, et al., 2005; Sjöström et al., 2007; Nadorff et al., 2011). Among those, Bernert *et al*. directly addressed the real question regarding research indicating that sleep disturbances may be specifically linked to suicidal behaviors: Is this link largely explained by depressive symptoms or how are specific symptoms of sleep disturbances relate to suicidal symptoms when controlling for depression? The 176 outpatients completed measures on sleep disturbances, suicidal symptoms. They controlled for depressive symptoms to establish a link between sleep disturbances and suicidality. They found that insomnia and nightmare symptoms were associated with both depressive symptoms and suicidality before controlling depressive symptoms. After controlling for depressive symptoms, only nightmares demonstrated an association with suicidal ideation. Another significant finding was that nightmares were particularly associated with suicidality among women compared to men. Before controlling for gender, a non-significant trend emerged between nightmare symptoms and suicidality, and this relationship remained after controlling for depression. After controlling for gender, the link between nightmare symptoms and suicidal ideation was statistically significant. This finding indicates that the association between nightmares and suicidality, while controlling for depression, was somewhat stronger among women versus among men (Bernert, et al., 2005).

For more than three decades, sleep disturbance had been considered the hallmark of posttraumatic stress disorder. Since insomnia has been observed in 90% of PTSD cases and

nightmare related to the trauma in 70%, this is understandable (Hendin et al., 2008). Besides, recently, Nadorff *et al.* clearly showed that nightmare symptoms may be "more than" a marker of PTSD, and hence may be important in the identification of suicidal ideation following several previous literatures (Nadorff et al., 2011).

8. Future remarks and treatment implications

8.1 Acute insomnia, the emergence of sleep psychiatry

Despite significant contributions made in the area of chronic insomnia, the area of acute insomnia has received comparatively little attention (Ellis et al., 2011). Overall, the findings from the review will highlight the need for a structured diagnosis of acute insomnia as the first step in a research and treatment strategy. Psychiatric and medical disorders are often associated with sleep disorders, especially acute insomnia which is a crucial element in clinical practices. Therefore, clinicians have to organize specific remedies for co-occurring acute insomnia itself. Recently the notion of sleep psychiatry (psychiatric therapeutic approach, both biologically and psychologically, based on sleep science) has gathered much attention worldwide (Goblin et al., 2004). Taking these points into account, when addressing potential treatment implications based on this conceptualisation of acute insomnia, three questions emerge: 1) Is it possible that acute insomnia can be identified and/or responded to in a timely manner?; 2) Is it possible that an intervention for acute insomnia has the potential to derail the occurrence of chronic insomnia?; 3) What would the optimal treatment approach be? Here, we will present a case study in order to show the possible efficiency of early sleep psychiatric intervention, mainly focusing on the subjective experience of the individual with acute insomnia (Abe et al., 2012).

8.1.1 Case study 2

Mr. T., a 25-year-old man, had a long history of OCD (Obsessive-Complusive Disorder) with recurrent obsessive thoughts of touching dirt and compulsive cleaning since his preadolescence. Firstly, at the age of 18, he consulted a psychiatrist for the purpose of treating his depressive symptoms after his father's sudden death. His depressive symptoms improved and then stabilized for several years with the aid of pharmacologic treatment (sulpiride 30mg, clorazepate 7.5mg and paroxetine 20mg). After graduating from professional school, he was able to work as a computer engineer in an urban company in spite of the persistence of his obsessive symptoms. One winter, he was addressed to our outpatient clinic by his general practitioner. His symptoms had already stabilized because of the same medication as a long-term maintenance treatment for OCD.

After four months of our follow-up, that spring, he was transferred to another section in his company. This change of social environment made him cogitate about his interpersonal relationship with other colleagues, which provoked acute insomnia symptoms, such as difficulty falling asleep and nighttime awaking. Additionally, he also suffered from daytime impairment related to his insomnia, especially hypersomnia and daytime sleepiness. He said, "I can't concentrate on my work because I have to fight to get to sleep" and "I feel afraid of falling asleep". Typically, the fear of insomnia was exacerbated. In other words, he was very afraid of losing his career position in his new section caused by the daytime impairments (e.g., losing concentration and diminished performance), which he attributed to his insomnia.

In order to improve acute insomnia symptoms, we treated him mainly with an early sleep psychiatric approach as a non-pharmacological intervention. Intentionally, we avoided increasing medication, because his principal concerns were strongly related with daytime impairment of insomnia. Adding another medication to improve sleep might run a risk of exacerbating daytime consequences of insomnia. In this situation, we treated him, making use of a home-monitoring actigraphy and an oxygen saturation tool. After two days of monitoring, the actigraphy always measured total sleep time and number of nighttime awakenings, so data showed that he had slept sufficiently during the night contrary to his complaints. As a first step, we supported him by showing him recorded sleep data, which suggested that his objective quality of sleep was different from his subjective complaining. This manifestation explained by itself a typical psychopathology of insomnia. This monitoring continued for a week. During this period, he was encouraged to visit our clinic 3 times a week for evaluation. Over nights, his actigraphic records always suggested longer and more consolidated sleep efficiency compared with his subjective intensity of insomnia complaints. As a next step, one night he was asked to get installed a portable oxygen saturation tool. The obtained data showed that some presence of hypoxemia during his sleep, which could partially explain the fragility of his sleep function. Also, he was found drinking alcohol and smoking just before going to bed. Moreover, he often surfed the web in eating snacks during the night on weekends. Therefore, we considered this above data as important evidence to strongly stop him from smoking and drinking before bedtime, and urging him to keep regular habits even at the weekend. With this intervention equipped with the home-monitored objective data, also based on sleep hygiene education (e.g. avoid bedside drinking, smoking, snacking and surfing the internet), his anxiety and fear of insomnia diminished dramatically and he spontaneously recovered from acute insomnia.

The lifetime prevalence of OCD is comparatively high at 2-3.5% of the population. While neither the core syndromal manifestations nor prominent associated features of OCD include sleep disturbances, patients suffering from OCD often complain about their sleep disturbance. Clinical observations show that their complaints are non-specific and persist. Previous sleep studies among patients with OCD are sparse and results inconsistent, often confounding with their comorbid depressive illness. Psychiatric disorders, such as neurotic disorders including OCD, are often associated with sleep disorders, especially insomnia which is a crucial element in clinical practices. Characteristically, insomniacs often complain about their sleep more than about the lack of it objectively measured. Therefore, we have to organize specific remedies for co-occurring acute insomnia itself. Possible other reasons for explaining his diminished quality of sleep in this case, were as follows; 1) presence of co-occurring subclinical depressive symptoms, 2) negative consequences of core OCD symptoms of sleep habits, 3) concurrent diurnal side effects of long-term prescribed medication. Clinically, these aspects must always be taken into consideration for managing sleep disturbance comorbid with neurotic disorders including OCD.

In this case, we attempted to have an early intervention in the vicious cycle of acute insomnia. This early sleep focused intervention prevented him from entering the chronic vicious cycle of psycho/physiological hyperarousal, which was supposed to play a central role in the pathophysiology of insomnia. We emphasize several suggestions about acute insomniac state. “I can’t sleep”, “I don’t get enough sleep”: This kind of complaints have often led to the easiest solution of direct prescriptions of hypnotics. The accumulation of

hypnotics has eventually had negative consequences in their everyday QOLs, such as daytime sleepiness and diminished concentration. Traditionally, Morita Therapy, a unique psychotherapy originated in Japan was devised for treating classical neurotic disorder. That concept has evolved the phenomenology of insomniacs as a subjective fabricated nature, claiming that clinicians are liable to make an error by just giving hypnotics to help the patient's feeling of sleeplessness without attempting radical cure on him. In this case, theoretically we applied some conventional concepts of Morita therapy to the treatment, utilizing the latest home monitoring instruments. We have to understand the fundamental phenomenology of diminished quality of sleep, and then give feedback to the acute insomniacs themselves in an appropriate way. To explain this process in the Morita theory, we attempted to stop exacerbating "psychic interaction" of acute insomniacs. This way of feedback may have something in common with the current well-developing Mindful-Based Cognitive Behavioral Therapy for insomnia. Despite their nature of subjective-objective discrepancy, individuals suffering from acute insomnia are situated under a subjectively perceived overwhelming threat.

We may stress that focusing on how sleep state misperception could be a particularly central aspect of insomnia within the context of OCD. From this case study, it appears that the treatment was largely successful because the actigraphic records helped to correct the patient's misperceptions. Perhaps, such a focused intervention has a non-specific and positive psychotherapeutic effect. This could also have implications about the possible application of actigraphy to treat sleep problems within anxiety disorders. A home monitoring system, such as actigraphy, not only provides satisfactory objective evaluation, but also a supportive psychotherapeutic effect in diminishing fear and anxiety related with acute insomnia. Getting an individual to recognize at an early stage, and providing him with treatment pathway guided by actigraph to deal with, were crucial in this case.

8.2 Alcoholism and insomnia

Some researchers have investigated between sleep disturbances in an adolescent period and its temporal development of substance abuse. The role of alcohol in the suicidal process warrants special attention (Gromov, I & Gromov, D, 2009; Wong et al., 2010). Adolescents may also be considered to choose drinking habits and substance abuse as a self-help sleep habit in order to escape negative suicide ideation. The association between insomnia and its self-medication with alcoholism has been acknowledged. The relationship between sleep problems and substance use/abuse has been demonstrated in both adults and adolescents. Insomnia has been shown to prospectively predict alcohol problems among some adults (Brower et al., 2001).

One recent unpublished data in alcoholic groups in a psychiatric hospital in Japan also showed that the majority of middle-aged alcoholic patients entering treatment reported insomnia symptoms and recognized themselves their diminished quality of sleep (Asami et al., 2011). In the Epidemiological Catchment Area study in the USA (n= 7,954), individuals with persistent insomnia at baseline were more likely than individuals without insomnia to experience the first onset of alcohol abuse or dependence one year later (Ford & Kamerow, 1989).

Recently, Pieters *et al*. have investigated the associations between pubertal development, sleep preference, sleep problems, and alcohol use in 431 early adolescents (Pieters et al., 2010). Then, they studied whether the associations changed when controlling for adolescent internalizing and externalizing problems. Results showed that pubertal development was positively associated with sleep problems and more evening-type tendencies (e.g., favouring later bedtimes), which in turn were positively related to alcohol use. From this study, it can be concluded that both puberty and sleep regulation are important factors in explaining alcohol use in early adolescence. This research has shown again a possible bi-directional relation between alcohol use and sleep, while profound puberty-dependent transitions regarding sleep patterns take place in early adolescence. Moreover, puberty has been associated with an increase in alcohol use of adolescents. They provided valuable data to understand the relationship among puberty, sleep problems, and alcohol use. Questions regarding that association, the possible reciprocal relationships among circadian phase preference, sleep problems and alcohol use, mediators and moderators of such relationships, as well as potential gender differences on these relationships were discussed (Wong, 2010). To understand the relationships among pubertal development, sleep problems, and alcohol use, researchers need to be aware of the physiological changes that take place in puberty, as well as the psychosocial factors that are associated with such changes (Pieters et al., 2010).

To the best of our knowledge, Wong et al. s' several reports in the USA have been the only longitudinal study examining the relationship between childhood sleep problems and adolescent substance use (Wong et al., 2004, 2009, 2010). Following their early works from a community sample of high-risk families and controls (292 boys and 94 girls), they have also tested whether adolescent sleep problems and poor response inhibition mediated the relationship between childhood sleep problems and substance (alcohol and drug) outcomes in young adulthood (Wong et al., 2010). Eventually, longitudinal design should be useful. Prevention and intervention programs may want to consider the role of sleep problems and response inhibition on substance use and abuse.

8.3 Insomnia and trauma in current Japanese society aftermath of tsunami disaster

A massive 9.0-magnitude earthquake occurred in the Pacific Ocean near Northeast Japan on March 11, 2011, causing serious damage to Japan. The effect of the East Japan Earthquake will not terminate within months. Many survivors experienced observing the disaster of the tsunami wiping out everything, and those extraordinary experiences will surely cause trauma among many minors (children and adolescents) who survived this disaster (Takeda, 2011). Those affected adults and adolescents should be taken care of closely for the possible occurrence of post-traumatic stress disorder, in terms of daily stress coping and sleep hygiene parameters related with insomnia. How does such emotional affects predict insomnia and nightmare symptoms? Bereavement is a ubiquitous part of the human condition. Almost no person makes it through his or her life without having to cope with the loss of a loved one several different times. The loss of a parent, child, or grandparent can be very distressing. From now on, further research will be needed to investigate the relationship between bereavement, complicated grief and recovery sleep.

Before concluding this discussion, we cited another provident remark, proposed by a French psychologist. She challenged to develop an analysis of various external and intrapsychic

factors related to adult insomnia (Dollander, 2002). The author suggests some preventive perspective to face the etiology of adult insomnia, and points out limits of pharmacological treatment. From a clinical viewpoint, she succeeded in making methodological propositions to test the four exposed psychodynamic assumptions: 1) insomnia should be a result of anguish excess related to intrapsychic conflicts which can't lead to a mental elaboration; 2) insomnia should be a result of psychic functioning invalidation; 3) some insomnias are related to waking associated with repetitive nightmares, and 4) some insomnias are in relation with an impossibility to accept passive position. These aspects are still in a hypothetical model, but it should always be noted in constructing long-term treatment strategy targeting on insomnia especially in the aftermath of trauma.

9. Brief summary and conclusion

In summary, clinical and epidemiological studies suggest that sleep disturbances (insomnia) are closely associated with suicidality and other correlates both in adult, and probably more importantly, in adolescent. In some cases, this association appears to exist above and beyond depression and PTSD. Nightmare shows a unique association with suicide risk, whereas additional research is needed to clarify both pathophysiology and symptomatology of insomnia. Studies should also be undertaken to determine the effects of adequate sleep and sleep hygiene promotion on mental health and suicide prevention. As the association between alcohol use and sleep has also been well documented among adults, researchers need to be aware of the pubertal development that takes place in sleep problems and their coping strategies. Circadian and homeostatic factors drive sleep. The sleep focused intervention studies may help us learn more about the fundamental role and homeostatic process of sleep dynamics in psychiatric disorder. Issues regarding the relationship between puberty and insomnia, the possible reciprocal relationships among circadian phase preference, sleep problems and alcohol use, mediators and moderators of such relationships (i.e. risk, protective and resilient factors), as well as potential gender differences on these relationships were discussed in further research and clinical practices.

10. Acknowledgment

This work was partially supported by a Health Science Research Grant from the Ministry of Helth, Labour and Welfare. The authors report no other financial affiliation or relationship relevant to the subject of this article. The views expressed in this article are mainly those of one of the authors (Y.A) and do not necessarily reflect the official policy or position of the authors' institutions.

11. References

[1] Abe, Y. & Mishima, K. (2008). The concepts and pathophysiology of insomnia. *Brain 21*, 11, 62-68, [Article in Japanese].

[2] Abe, Y.; Mishima, K.; Kaneita, Y.; Li, L.; Ohida, T.; Nishikawa, T. & Uchiyama, M. (2011). Stress coping behaviors and sleep hygiene practices in a sample of Japanese adults with insomnia. *Sleep and Biological Rhythm*, 9, 35–45.

[3] Abe, Y.; Nishimura, G. & Endo, T. Early sleep psychiatric intervention for acute insomnia: Implications from a case of Obsessive-Compulsive Disorder (2012). *J. Clin. Sleep Med*., in press.

[4] Ağargün, M.Y.; Kara, H. & Solmaz, M. (1997). Subjective sleep quality and suicidality in patients with majour depression. *J. Psychiatr. Res.*, 31, 377-381.

[5] Ağargün, M.Y.; Cilli, A.S.; Kara, H.; Tarhan, N.; Kincir, F. & Oz, H. (1998). Repetitive and frightening dreams and suicidal behavior in patients with major depression. *Compr. Psychiatry*, 39, 198-202.

[6] Ağargün, M.Y.; Kara, H.; Ozer, O.A..; Selvi, Y.; Kiran, U. & Kiran, S. (2003). Nightmares and dissociative experiences: The key role of childhood traumatic events. *Psychiatry and Clin. Neurosciences*, 57, 139–145.

[7] Ağargün, M.Y. & Beşiroğlu, L. (2005). Sleep and suicidality: do sleep disturbances predict suicide risk? *Sleep*, 28, 1039-1040.

[8] AASM (American Academy of Sleep Medicine). (2005). *International Classification of Sleep Disorders, 2nd Ed.: Diagnostic and Coding Manual*. American Academy of Sleep Medicine (ICSD-2). Westchester, IL.

[9] American Psychiatric Association (APA). (2000). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Text revision. American Psychiatric Association: Washington, D.C.

[10] Apter, A.; Gothelf, D.; Orbach, I.; Weizman, R.; Ratzoni, G.; Har-Even, D. & Tyano, S. (1995). Correlation of suicidal and violent behavior in different diagnostic categories in hospitalized adolescent patients. *J. Am. Acad. Child Adolesc. Psychiatry*, 34, 912–918.

[11] Asami, M.; Abe, Y.; Suzuki, R.; Hasuo, R.; Nirasawa, H.; Jukuroki, H. & Kakibuchi, Y. (2011). A study on the subjective sleep evaluation and the related factors in the alcoholic, presented at the 36th Annual Metting of Japanese Society of Sleep Research, Oct 16.

[12] Bader, K.; Schäfer, V.; Schenkel, M.; Nissen, L. & Schwander, J. (2007). Adverse childhood experiences associated with sleep in primary insomnia. *J. Sleep Res.*, 16, 285-296.

[13] Basta, M.; Chrousos, G.P.; Vela-Bueno, A. & Vgontzas, A.N. (2007). Chronic Insomnia and Stress System. *Sleep Med. Clin.*, 2, 279-91.

[14] Baglioni, C.; Spiegelhalder, K.; Lombardo, C. & Riemann, D. (2010). Sleep and emotions: a focus on insomnia. *Sleep Med. Rev.*, 14, 4, 227-238.

[15] Bastien, C. & Morin, C. M. (2000). Familial incidence of insomnia. *J. Sleep Res.*, 9, 49–54.

[16] Bastien, C. H.; Vallières, A. & Morin, C. M. (2004). Precipitating factors of insomnia. *Behav. Sleep Med,* 2, 50–62.

[17] Beaulieu-Bonneau, S.; LeBlanc, M.; Merette, C.; Dauvilliers, Y. & Morin, C. (2007). Family history of insomnia in a population-based sample. *Sleep,* 30, 1739–1745.

[18] Bernert, R.A.; Joiner, T.E. Jr.; Cukrowicz, KC; Schmidt, N.B. & Krakow B. (2005). Suicidality and sleep disturbances. *Sleep*, 28, 1135-1141.

[19] Bernert, R. A. & Joiner, T.E. (2007). Sleep disturbances and suicide risk: A review of the literature. *Neuropsychiatr. Dis. Treat.*, 3, 735-743.

[20] Blenkiron, P; House, A. & Milnes, D. (2000). The timing of acts of deliberate self-harm: is there any relation with suicidal intent, mental disorder or psychiatric management? *J. Psychosom. Res,* 49, 3-6.

[21] Bonnet, M.H. & Arand, D.L. (1995). 24-Hour metabolic rate in insomniacs and matched normal sleepers. *Sleep,* 18, 581–588.

[22] Bonnet M.H. & Arand, D.L. (2010). Hyperarousal and insomnia: state of the science. *Sleep Med. Rev.,* 14, 9-15.

[23] Brand, S.; Luethi, M.; von Planta, A; Hatzinger, M. & Holsboer-Trachsler, E. (2007). Romantic love, hypomania, and sleep pattern in adolescents. *J. Adolesc. Health.*, 41, 1, 69-76.

[24] Brand, S. & Kirov, R. (2011). Sleep and its importance in adolescence and in common adolescent somatic and psychiatric conditions. *Int. J. Gen. Med,* 4, 425-442.

[25] Bridge, J.A.; Goldstein, T.R. & Brent, D.A. (2006). Adolescent suicide and suicidal behavior. *Journal of Child Psychology and Psychiatry,* 47, 372-394.

[26] Brower, K.J.; Aldrich, M.S.; Robinson, E.A.; Zucker, R.A. & Greden, J.F. (2001). Insomnia, self-medication, and relapse to alcoholism. *Am. J. Psychiatry,* 158, 399-404.

[27] Buysse, D.J. (2008). Chronic insomnia. *Am. J. Psychiatry*, 165, 678–686.

[28] Buysse, DJ; Angst, J; Gamma, A; Ajdacic, V; Eich, D. & Rössler, W. (2008). Prevalence, course, and comorbidity of insomnia and depression in young adults. *Sleep,* 31, 473–480.

[29] Choquet, M. & Menke, H. (1990). Suicidal thoughts during early adolescence: prevalence, associated troubles and help-seeking behavior. *Acta Psychiatr. Scand.,* 81, 170-177.

[30] Choquet, M.; Kovess, V. & Poutignat, N. (1993). Suicidal thoughts among adolescents: an intercultural approach. *Adolescence,* 28, 111, 649-659.

[31] Csóka, S.; Simor, P.; Szabó, G.; Kopp, M.S. & Bódizs, R. (2011). Early maternal separation, nightmares, and bad dreams: results from the Hungarostudy Epidemiological Panel. *Attach Hum Dev.*, 13, 125-140.

[32] Dauvilliers, Y.; Morin, C.; Cervena, K.; Carlander, B.; Touchon, J.; Besset, A. & Billiard, M. (2005). Family studies in insomnia. *J. Psychosom. Res.*, 58, 271–278.

[33] Dollander, M. (2002). Etiology of adult insomnia. *L'Encéphale*, 28, 493-502 [Article in French].

[34] Drake, C.L.; Scofield, H. & Roth, T. (2008). Vulnerability to insomnia: the role of familial aggregation. *Sleep Med.*, 9, 297–302.

[35] Edinger, J.D., Means, M.; Carney, C. E. & Krystal, A.D. (2008). Psychomotor performance deficits and their relation to prior nights' sleep among individuals with primary insomnia. *Sleep*, 31, 599–607.

[36] Ellis, J.G.; Gehrman, P.; Espie, C.A.; Riemann, D. & Perlis, M.L. (2011). Acute insomnia: Current conceptualizations and future directions. *Sleep Med. Rev.,* doi:10.1016/j.smr. 2011.02.002.

[37] Endo, S. (1962). The Psychophysiological Study of Neurotic insomnia. *Psychiatria et Neurologia Japonica*, 64, 673-707.

[38] Fawcett, J.; Scheftner, W.A.; Fogg, L.; Clark, D.C.; Young, M.A.; Hedeker, D. & Gibbons, R. (1990). Time-related predictors of suicide in major affective disorder. *Am. J. Psychiatry,* 147, 1189-1194.

[39] Ford, D. E. & Kamerow, D. B. (1989). Epidemiologic study of sleep disturbances and psychiatric disorders: an opportunity for prevention? *JAMA,* 262: 1479–1484.

[40] Franzen, P.L.& Buysse, D.J. (2008). Sleep disturbances and depression : risk relationships for sebsequent depression and therapeutic implications. *Dialogues Clin. Neurosci.*, 10, 473-481.

[41] Fujino, Y.; Mizoue, T; Tokui, N. & Yoshimura, T. (2005). Prospective cohort study of stress, life satisfaction, self-rated health, insomnia, and suicide death in Japan. *Suicide Life-Threat Behav.*, 35, 2, 227–237.

[42] Gangwisch, J.E.; Babiss, L.A.; Malaspina, D.; Turner, J.B.; Zammit, G.K. & Posner, K. (2010). Earlier parental set bedtimes as a protective factor against depression and suicidal ideation. *Sleep*, 33, 97-106.

[43] Gau, S. F. & Soong, W.T. (1995). Sleep problems of junior high school students in Taipei. *Sleep,* 18, 667–673.

[44] Germain. A.; Shear, K.; Monk, T.H.; Houck, P.R.; Reynolds, C.F.; Frank, E. & Buysse, D.J. (2006). Treating complicated grief: effects on sleep quality. *Behav. Sleep Med.*, 4, 3, 152-163.

[45] Germain, A.; Buysse, D.J. & Nofzinger, E. (2008). Sleep-specific mechanisms underlying posttraumatic stress disorder: integrative review and neurobiological hypotheses. *Sleep Med. Rev.*, 12, 3, 185-195.

[46] Germain, A.; Hall, M.; Krakow, B.; Shear, K. M. & Buysse, D.J. (2005). A brief sleep scale for Posttraumatic Stress Disorder: Pittsburgh Sleep Quality Index Addendum for PTSD. *J. Anxiety Disord.*, 19, 2, 233-244.

[47] Goldstein, T. R.; Bridge, J.A. & Brent, D.A. (2008). Sleep disturbance preceding completed suicide in adolescents. *J. Consult. Clin. Psychol.*, 76, 84-91.

[48] Goblin, A.; Kravitz, H. & Keith, L. eds. (2004). Sleep Psychiatry. Taylor & Francis, London.

[49] Gregory, A.M.; Caspi, A.; Moffitt, T.E. & Poulton R. (2006). Family conflict in childhood: a predictor of later insomnia. *Sleep,* 29, 1063-1067.

[50] Gromov, I. & Gromov, D. (2009). Sleep and substance use and abuse in adolescents. *Child. Adolesc. Psychiatr. Clin. N. Am.*, 18, 929-946.

[51] Hasler, B. & Germain, A. (2009). Correlates and Treatments of Nightmares in Adults. *Sleep Med Clin.,* 4, 507-517.

[52] Hendin, H.; Maltsberger, J.T. & Szanto, K. (2008). The psychosocial context of trauma in treating PTSD patients. *Am. J. Psychiatry,* 165, 28-32.

[53] Jefferson, C.D.; Drake, C.L.; Scofield, H.M., Myers, E.; McClure, T.; Roehrs, T. & Roth, T. (2005). Sleep hygiene practices in a population-based sample of insomniacs. *Sleep,* 28, 611–615.

[54] Johnson, E.O.; Roth, T.; Schultz, L. & Breslau, N. (2006). Epidemiology of DSM-IV insomnia in adolescence: lifetime prevalence, chronicity, and an emergent gender difference. *Pediatrics,* 117, 247-256.

[55] Kaneita, Y.; Ohida, T.; Osaki,Y.; Tanihata, T.; Minowa, M.; Suzuki, K.; Wada, K.; Kanda, H. & Hayashi, K. (2006). Insomnia among Japanese adolescents: anationwide representative survey. *Sleep,* 29, 1543–1550.

[56] Krakow, B. ; Artar, A. & Warner, T.D. (2000). Sleep disorder, depression and suicidality in female sexual assault survivors. *Crisis,* 21, 163-170.

[57] Krakow, B.; Ribeiro, J.D.; Ulibarri, V.A.; Krakow, J. & Joiner, T.E. Jr. (2011). Sleep disturbances and suicidal ideation in sleep medical center patients. *J. Affect. Disord.,* 131, 422-427.

[58] LeBlanc, M.; Beaulieu-Bonneau, S.; Mérette, C.; Savard, J.; Ivers, H. & Morin, C.M. (2007). Psychological and health-related quality of life factors associated with insomnia in a population-based sample. *J. Psychosom. Res.,* 63, 157–166.

[59] Levin, R. & Nielsen, T.A. (2007). Disturbed dreaming, posttraumatic stress disorder, and affect distress: A review and neurocognitive model. *Psychological Bulletin,* 133, 482–528.

[60] Levin, R. & Nielsen, T. (2009). Nightmares, bad dreams and emotion dysregulation: A review and new neurocognitive model of dreaming. *Current Directions in Psychological Science,* 18, 84–88.

[61] Li, S.X., Lam, S.P.; Yu, M.W.; Zhang, J. & Wing, Y.K. (2010). Nocturnal sleep disturbances as a predictor of suicide attempts among psychiatric outpatients: a clinical, epidemiologic, prospective study. *J. Clin. Psychiatry*, 71, 1440-1446.

[62] Liu, X.C.; Uchiyama, M.; Okawa, M. & Kurita, H. (2000). Prevalence and correlates of self-reported sleep problems among Chinese adolescents. *Sleep,* 23, 27–34.

[63] Liu, X. & Buysse, D.J. (2006). Sleep and youth suicidal behavior: a neglected field. *Curr. Opin. Psychiatry,* 19, 288-293.

[64] Mahgoub, N. A. (2009). Insomnia and suicide risk. *J. Neuropsychiatry Clin. Neurosci,* 2, 21, 232-233.

[65] Mai, E. & Buysse, D.J. (2008). Insomnia: Prevalence, Impact, Pathogenesis, Differential Diagnosis, and Evaluation. *Sleep Med. Clin.*, 3, 2, 167-174.

[66] Morin, C.M.; Rodrigue, S. & Ivers, H. (2003). Role of stress, arousal, and coping skills in primary insomnia. *Psychosom. Med.,* 65, 259–267.

[67] Morin, C.M. & Espie, C. (2004). Insomnia: A Clinical Guide to Assessment and Treatment. Springer: New York.

[68] Morin, C.M.; Leblanc, M.; Daley, M.; Gregoire, J.P. & Merette, C. (2006). Epidemiology of insomnia: prevalence, self-help treatments, consultations, and determinants of help-seeking behaviors. *Sleep Med,* 7, 123–130.

[69] Morrison, D.N.; McGee, R. & Stanton, W.R. (1992). Sleep problems in adolescence. *J. Am. Acad. Child. Adolesc. Psychiatry,* 31, 94–99.

[70] Motohashi, Y. (1990). Circadian variation in suicide attempts in Tokyo from 1978 to 1985. *Suicide Life Threat. Behav.,* 20, 533-539.

[71] Munezawa, T.; Kaneita, Y.; Osaki, Y.; Kanda, H.; Ohtsu, T.; Suzuki, H.; Minowa, M.; Suzuki, K.; Higuchi, S.; Mori, J. & Ohida, T. (2011). Nightmare and sleep paralysis among Japanese adolescents: a nationwide representative survey. *Sleep Med.*, 12, 56-64.

[72] Nadorff, M. R.; Nazem, S. & Fiske, A. (2011). Insomnia symptoms, nightmares, and suicidal ideation in a college student sample. *Sleep,* 34, 93-98.

[73] National Institutes of Health (NIH). (2005). National Institutes of Health state of the science conference statement on manifestations and management of chronic insomnia in adults. June 13–15, 2005, *Sleep,* 28, 1049–1057.

[74] Nielsen, T.A.; Stenstrom, P. & Levin, R. (2006). Nightmare frequency as a function of age, gender and September 11, 2001: Findings from an Internet questionnaire. *Dreaming,* 16, 145–158.

[75] Nir, Y. & Tononi, G. (2010). Dreaming and the brain: from phenomenology to neurophysiology. *Trends Cogn. Sci.,* 14, 2, 88-100.

[76] Nofzinger, E.A.; Buysse, D.J.; Germain, A.; Price, J.C.; Miewald, J.M. & Kupfer, D.J. (2004). Functional neuroimaging evidence for hyperarousal in insomnia. *Am. J. Psychiatr*, 161, 11, 2126-2128.

[77] Ohayon, M.M. (2002). Epidemiology of insomnia: what we know and what we still need to learn. *Sleep Med. Rev.*, 6, 97–111.

[78] Ohayon, M. M. & Lemoine, P. (2004). Daytime consequences of insomnia complaints in the French general population. L'Encéphale, 222–227 [Article in French].

[79] Ohida, T.; Osaki, Y.; Doi, Y.; Tanihata, T.; Minowa, M.; Suzuki, K.; Wada, K.; Suzuki, K. & Kaneita, Y. (2004). An epidemiological study of self-reported sleepproblems among Japanese adolescents. *Sleep,* 27, 978–985.

[80] Orff, H.J.; Drummond, S.P.; Nowakowski, S. & Perlis, M.L. (2007). Discrepancy between subjective symptomatology and objective neuropsychological performance in insomnia. *Sleep,* 30, 1205-1211.

[81] Pandey, G. N. (2011). Neurobiology of adult and teenage suicide. *Asian Journal of Psychiatry*, 4, 2-13.

[82] Pieters, S.; Van der Vorst, H.; Burk, W.J.; Wiers, R.W. & Engels, R.C. (2010). Puberty-dependent sleep regulation and alcohol use in early adolescents. *Alcohol Clin. Exp. Res.*, 34, 1512-1518.

[83] Pigeon, W.R. & Caine, E.D. (2010). Insomnia and the risk for suicide: does sleep medicine have interventions that can make a difference? *Sleep Med.*, 11, 9, 816-817.

[84] Proposed Draft Revisions to DSM-5 Disorders and Criteria. (2010). http://www.dsm5.org/.

[85] Pronger, C.E. (1914). Insomnia and suicide. Lancet, 184: 1356–1359.

[86] Riemann, D. & Voderholzer, U. (2003). Primary insomnia: a risk factor to develop depression? *J. Affect. Disord.* 76, 255–259.

[87] Riemann, D.; Voderholzer, U.; Spiegelhalder, K.; Hornyak, M.; Buysse, D.J.; Nissen, C. ; Hennig, J.; Perlis, M.L.; van Elst, L.T. & Feige, B. (2007). Chronic insomnia and MRI-measured hippocampal volumes: a pilot study. *Sleep*, 30, 955–958.

[88] Riemann, D.; Kloepfer, C. & Berger, M. (2009). Functional and structural brain alterations in insomnia: implications for pathophysiology. *Eur. J. Neurosci.*, 29, 9, 1754-1760.

[89] Riemann, D.; Spiegelhalder, K.; Feige, B.; Voderholzer, U.; Berger, M.; Perlis, M. & Nissen, C. (2010). The hyperarousal model of insomnia: a review of the concept and its evidence. *Sleep Med Rev.* 14, 19-31.

[90] Riemann, D.; Spiegelhalder, K.; Espie, C.; Pollmächer, T. ; Léger, D.; Bassetti, C. & van Someren, E. (2011). Chronic insomnia: clinical and research challenges--an agenda. *Pharmacopsychiatry*. 44, 1-14.

[91] Roane B.M. & Taylor, D.J. (2008). Adolescent insomnia as a risk factor for early adult depression and substance abuse. *Sleep,* 31, 1351-1356.

[92] Roberts, R. E.; Roberts, C.R. & Chen, I. G. (2001). Functioning of adolescents with symptoms of disturbed sleep. *Journal of Youth and Adolescence*, 30, 1-18.

[93] Rosen, J.; Reynolds, C.F.3rd.; Yeager, A.L.; Houck, P.R. & Hurwitz, L.F. (1991). Sleep disturbances in survivors of the Nazi Holocaust. *Am. J. Psychiatry*, 148, 62-66.

[94] Safer, D.J. (1997). Adolescent/adult differences in suicidal behavior and outcome. *Ann. Clin. Psychiatry,* 9, 61-66.

[95] Selvi, Y.; Aydin, A.; Boysan, M., Atli, A.; Agargun, M.Y. & Besiroglu, L. (2010). Associations between chronotype, sleep quality, suicidality, and depressive symptoms in patients with major depression and healthy controls. *Chronobiol Int.*, 27, 1813-1828.

[96] Shekleton, J. A.; Rogers, N. L. & Rajaratnam, S.M. (2010). Searching for the daytime impairments of primary insomnia. *Sleep Med. Rev.*, 14, 1, 47-60.

[97] Sher, L. (2008). Sleep disturbances, synaptic homeostasis and suicidal behaviour. *Aust. N. Z. J. Psychiatry,* 2008, 42, 1072-1073.

[98] Singareddy, R. K. & Balon, R. (2001). Sleep and suicide in psychiatric patients. *Ann. Clin. Psychiatry,* 13, 93-101.

[99] Sjöström, N.; Waern, M. & Hetta, J. (2007). Nightmares and sleep disturbances in relation to suicidality in suicide attempters. *Sleep,* 30, 91-95.

[100] Soldatos, C.R.; Allaert, F.A.; Ohta, T. & Dikeos, D. G. (2005). How do individuals sleep around the world? Results from a single-day survey in ten contries. *Sleep Med.,* 6, 5–13.

[101] Sompo Japan Research Institute Inc. Disease Management Reporter in Japan No. 17, May, 2010.

[102] Spielman, A. J.; Caruso, L.S. & Glovinsky, P.B. (1987). A behavioral perspective on insomnia treatment. *Psychiatr. Clin. North Am.*, 10, 541–553.

[103] Tamas, Z.; Kovacs, M.; Gentzler, A.L.; Tepper, P.; Gadoros, J.; Kiss, E.; Kapornai, K. & Vetró, A. (2007). The relations of temperament and emotion self-regulation with suicidal behaviors in a clinical sample of depressed children in Hungary. *J. Abnorm. Child Psychol.,* 35, 640-652.

[104] Taylor, D. J. & Roane, B.M. (2010). Treatment of insomnia in adults and children: a practice-friendly review of research. *J. Clin. Psychol.*, 66, 1137-1147.

[105] Taylor, D. J.; Gardner, C.E; Bramoweth, A.D; Williams, J.M.; Roane, B. M.; Grieser, E. A. & Tatum, J. I. (2011). Insomnia and Mental Health in College Students. *Behav. Sleep Med.*, 9, 107-116.

[106] Tononi, G. & Cirelli, C. (2003). Sleep and synaptic homeostasis: a hypothesis. *Brain Res. Bull.,* 62, 2, 143-150.

[107] Tononi, G. & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Med. Rev.,* 10, 49-62.

[108] Tsuno, N.; Besset, A. & Ritchie, K. (2005). Sleep and depression. *J. Clin. Psychiatry*, 66, 1254–1269.

[109] Takeda, M. (2011). Mental health care and East Japan Great Earthquake. *Psychiatry Clin. Neurosciences,* 65, 207-212.

[110] Tanskanen, A.; Tuomilehto, J.; Viinamäki, H.; Vartiainen, E.; Lehtonen, J. & Puska, P. (2001). Nightmares as predictors of suicide. *Sleep,* 24, 844-847.

[111] Troxel, W.M. & Germain, A. (2011). Insecure attachment is an independent correlate of objective sleep disturbances in military veterans. *Sleep Med.*, doi:10.1016/j.sleep.2011.07.005.

[112] Turvey, C.L.; Conwell, Y.; Jones, M.P.; Phillips, C.; Simonsick, E.; Pearson J.L. & Wallace, R. (2002). Risk factors for late-life suicide: a prospective, community-based study. *Am. J. Geriatr. Psychiatry,* 10, 398-406.

[113] Vollath, M.; Wicki, W. & Angst, J. (1989). The Zurich Study VIII. Insomnia: association with depression, anxiety, somatic symptoms, and course of insomnia. *Eur. Arch. Psychiatry Neurol. Sci.,* 239, 113–124.

[114] Wirz-Justice, A.; Benedetti, F. & Terman, M. (2009). Chronotherapeutics for Affective Disorders: A Clinician's Manual for Light and Wake Therapy. Basel, Karger.

[115] Wojnar, M.; Ilgen, M.A.; Wojnar, J.; McCammon, R.J.; Valenstein, M. & Brower, K.J. (2009). Sleep problems and suicidality in the National Comorbidity Survey Replication. *J. Psychiatr. Res.*, 43, 526-531.

[116] Wong, M.M; Brower, K.J.; Fitzgerald, H.E. & Zucker, R.A. (2004). Sleep problems in early childhood and early onset of alcohol and other drug use in adolescence. *Alcohol. Clin. Exp. Res.*, 28, 578-587.

[117] Wong, M. M.; Brower, K.J. & Zucker, R.A. (2009). Childhood sleep problems, early onset of substance use and behavioral problems in adolescence. *Sleep Med.,* 10, 787-796.

[118] Wong, M. M.; Brower, K. J.; Nigg, J.T. & Zucker, R.A. (2010a). Childhood sleep problems, response inhibition, and alcohol and drug outcomes in adolescence and young adulthood. *Alcohol. Clin. Exp. Res.,* 34, 1033-1044.

[119] Wong, M, M. (2010b). Pubertal development, sleep problems, and alcohol use: a commentary. *Alcohol. Clin. Exp. Res.*, 34, 2019-2021.

[120] Wong, M.M.; Brower, K.J. & Zucker, R.A. (2011). Sleep problems, suicidal ideation, and self-harm behaviors in adolescence. *J. Psychiatr. Res.,* 45, 505-511.

[121] Yang, C.M.; Wu, C.H.; Hsieh, M. H.; Liu, M.H. & Lu, F.H. (2003). Coping with sleep disturbances among young adults: a survey of first-year college students in Taiwan. *Behav. Med.*, 29, 133-138.

20

Saving More than Lives: A Gendered Analysis of the Importance of Fertility Preservation for Cancer Patients

Lisa Campo-Engelstein[1], Sarah Rodriguez[2] and Shauna Gardino[2,3]
[1]*Alden March Bioethics Institute & Department of OBGYN, Albany Medical College,*
[2]*Northwestern University, Feinberg School of Medicine,*
[3]*Oncofertility Consortium*
USA

1. Introduction

Cancer affects millions of Americans annually. Men's lifetime risk of developing cancer for all sites is 50%; women's lifetime risk is just over 33% (American Cancer Society, 2009). While cancer is generally perceived as a condition affecting people past their child-bearing years, nearly 10% of those diagnosed are under age 45 (Horner et al., 2009). Indeed, some of those diagnosed with cancer are still children. In 2006, an estimated 9,500 new cases of pediatric cancer were diagnosed in the United States (American Cancer Society, 2006). Because of recent breakthroughs and more aggressive treatments, the survival rate of those diagnosed with childhood cancer has risen to almost 80% (Clayman, Galvin, and Arnston, 2007). One estimate is that by 2010 one of every 250 adults will be a survivor of childhood cancer (Kinahan, 2007).

But while more aggressive treatments have meant more people survive cancer, these treatments have also resulted in impaired fertility or sterility for some. Given the numbers of children and adults within their child-bearing years diagnosed with, treated for, and surviving cancer, fertility concerns have emerged as a quality of life issue important to cancer survivors and their families. In one study of cancer survivors, 76% of those who were childless expressed a desire to have children in the future (Schover, 2009). Impaired fertility as a result of cancer treatment has physical as well as psychological effects. The existing literature on women whose fertility was impaired as a result of cancer treatment reveals an intense psychological distress; for these women, "psychological distress may result from, not only the loss of the physical ability to conceive, but also a symbolic loss of the option or idea of fertility, regardless of whether this would have been acted upon or achievable" (Carter et al, 2005, p. 93). Some studies on men have revealed similar levels of long-term distress over their impaired fertility as a result of cancer treatments (Schover, 2009) while other studies found that infertility is not nearly as devastating for men as it is for women.

Though reproduction is valued by both women and men, as the conflicting responses to studies between women and men (and even among men) illustrate, there are often

differences in how women and men respond to infertility. Gender has been the focus of much analysis in ethics, but little work has been done looking at gendered narratives to analyze fertility concerns as a quality of life issue among men and women undergoing possible fertility-impairing treatment for cancer. Such an analysis is relevant in the current context because of the rise in the number of younger people being diagnosed with and surviving cancer and accompanying scientific and technological advances in fertility preservation techniques. In the context of cancer survivorship, fertility has become a distinct quality of life consideration, with an entire new field (oncofertility) now dedicated to promoting fertility preservation options for cancer patients and survivors. Since oncofertility is a novel discipline that bridges a variety of academic scholarships, existing conceptual frameworks are inadequate for analyses within this field. We propose a multidisciplinary approach to understand the gendered themes of infertility as a quality of life concern in the current context of oncofertility.

In doing so, we argue that exploring common cultural conceptions of gender is essential to understanding cancer patients' and their families' fertility preservation decision making when confronting potential infertility. Our analysis begins at the patient level with a discussion of the existing literature that describes gender differences among cancer patients regarding fertility, cancer treatment, and the effects of both on their quality of life. Uniquely, we contextualize these existing social science studies within a historical context, using a new perspective to attempt to explain the sometimes divergent responses between women and men and among men by focusing on the different ways men and women have been treated for infertility broadly since the nineteenth century. We build on this historical framework by conducting a gendered analysis of infertility and fertility choices that is centered in the current cultural climate. In this way, we address gender, infertility, and fertility choices from three distinct levels—the individual level, the contextualizing historical level, and the broader current cultural level—creating our own multidisciplinary framework. Our analysis offers an insightful approach, creating a new framework within which we are able to draw meaningful conclusions about the impact gendered responses to infertility have on cancer patients' medical care and how health care providers and researchers can incorporate this knowledge to improve patient care.

2. Fertility preservation technologies

The developing field of oncofertility is dedicated to providing fertility preservation options to cancer patients, and a number of alternatives currently exist for both men and women. Reproductively mature men confronting a cancer diagnosis can generate a sperm sample and cryopreserve (freeze) their gametes for later use. Ejaculation can be stimulated in young men and those too sick to produce a sperm sample, or an experimental testicular biopsy can be done; the resulting sample can be cryopreserved. The technology for freezing sperm is well-established and successful, leaving these men with a viable option to become a biological father after cancer. Freezing sperm is relatively inexpensive and easy to accomplish, thus a feasible option for men of all socioeconomic backgrounds and cancer types.

Women's reproductive potential can also be jeopardized during cancer treatment. The only established method of fertility preservation for women is embryo banking, but this is often not a palatable option for young and/or single women. Egg banking is gaining popularity because there is no reliance on a male donor (known or anonymous); however, it is still

considered an experimental intervention by the American Society for Reproductive Medicine (ASRM). Neither embryo nor egg banking, however, are not good options for some newly diagnosed cancer patients because these procedures require a two-to-three week delay in cancer treatment and cannot be performed on those who have not yet reached puberty. For women or young girls for whom embryo or egg banking may not be an acceptable option, ovarian tissue cryopreservation offers another opportunity to protect their potential to parent. Ovarian tissue cryopreservation is a procedure in which one ovary is removed and ovarian tissue is frozen in small strips. Based on the woman's treatment plan, the strips can be transplanted back and potentially restore ovarian function. Researchers are also currently working on another way to use the tissue: maturing the follicles from the cryopreserved ovarian tissue within the laboratory. As the functional unit of the ovary, the follicles would ideally mature into eggs and then be fertilized, with the resulting embryo implanted back into the woman when she desires pregnancy (Jeruss and Woodruff, 2009).

3. Examining the existing research on QOL and fertility

As we will discuss in-depth in the next section, infertility has historically been associated with women. Indeed, early studies of fertility concerns among cancer patients and survivors claim that women value fertility to a greater extent than men. In 1987, Wasserman, Thomspon, Wilimas, and Fairclough (1987) determined that attitudes toward possible sterility differed dramatically between female and male Hodgkin's survivors, with females expressing much more concern about their childbearing potential than males. Five years later, Zelter published a literature review, concluding that women as a group seek more information and evaluation of their fertility status than men (Zelter, 1993). Schover, Rubicki, Martin, and Bringelsen (1999) followed up with another literature review, hypothesizing from gathered data that women are more distressed over infertility, more concerned about having children, and more likely to see parenthood as an integral part of their life goals when compared to men. Finally, Patridge et al (2004) found that women will even sacrifice the efficacy of cancer treatment to lessen their chances of infertility or sterility, describing how, if given a choice, young women with early-stage breast cancer may choose a less toxic regimen of chemotherapy even if it confers slightly less protection from recurrence of cancer.

However, a handful of studies have recognized that male cancer patients and survivors value their fertility as well. A 1990 study by Reiker et al of 153 testicular cancer survivors indicated that distress about infertility is also prevalent among men, particularly among those who have cancer treatments that are likely to severely impair fertility (Reiker, 1990). Similarly, a 2003 study by Green, Galvin, and Horne confirmed that infertility can cause long-term distress among men with cancer. The literature regarding fertility concerns among male cancer patients is scarce when compared to the number of studies that demonstrate the importance of fertility among female cancer patients, but nonetheless this data should not be ignored. A growing body of more recent literature is beginning to recognize fertility concerns among male cancer patients, concluding that gender differences in fertility concerns may not be as prominent as once thought.

Indeed, a 1999 study by Schover et al, which used a questionnaire to examine 283 young cancer survivors, found that about 80% of cancer survivors viewed themselves positively as

actual or potential parents, with no observed gender-related differences in the wish to have children or distress about fertility (Schover, et al, 1999). Similarly, a 2004 study by Zebrack, Casillas, Nohr, Adams, and Zelter used semi-structured interviews to assess the impact of cancer on long-term cancer survivors' quality of life, concluding that both men and women expressed a desire to have children in the future (Zebrack, 2004). Finally, an exploratory qualitative study by Crawshaw (2010) found that fertility matters affected identity, well-being and life planning as well as reproductive function, regardless of gender.

Our literature review of gender differences in fertility concerns among cancer patients indicates that, at present, there is conflicting evidence, but understandings of these gendered differences may be changing. While older studies indicate that female cancer patients value their fertility more strongly than male cancer patients, newer research suggests that these attitudes may be changing as more male cancer patients are beginning to recognize and express their fertility-related distress. However, these disparate conclusions could also be explained by differing methodologies; the aforementioned studies range from semi-structured interviews to questionnaires to literature reviews. As these studies are rooted in distinct and differing methodologies, drawing accurate comparisons between them may be complicated.

In the rest of the paper we explore factors that may account for the aforementioned discrepancies in the literature relating to fertility concerns among female and male cancer patients, using both historical and cultural frameworks to explain how these attitudes are developed and derived.

4. Historical foundations: Women and infertility

In order to better understand the basis for the differing ways women and men may view infertility today, and the lack of consensus about the value of fertility among men, we need to explore the differences in the ways men and women have been medically treated for infertility in this country. As we will next describe, medical treatment for infertility in the United States reinforced and reflected prevailing cultural ideas about masculinity and femininity.

Though Americans could consult medical guides in the late eighteenth century for recourse to alleviate an involuntarily childless marriage, it was during the nineteenth century when physicians became increasingly involved in treating, and patients began to increasingly seek medical intervention for, infertility (Marsh and Ronner, 1996). Almost always, however, the patients seeking treatment were women. Why? Though motherhood had earlier been a principal role for women, during the nineteenth century it increasingly became the defining role for women, especially white middle-class women (Apple, 1997; Marsh and Ronner, 1996). By the early nineteenth century, American society began to draw clearer lines between family and community. These lines changed how families were seen and composed, both of which had profound implications for women's conceptions of femininity and reproduction. During the course of the nineteenth century, biological parenthood rather than household composition came to define a family, and thus being a mother increasingly meant bearing one's own children. As part of this shift, biological motherhood was increasingly regarded as the primary and principal role of women – in contrast to women in colonial America, where motherhood, though important, was seen as only one of a woman's

roles (Marsh and Ronner, 1996). Being a mother, especially through pregnancy, began to be more strongly tied to being a woman. This change in the stress placed on biological motherhood as the core identifier for femininity, along with the rise in the profession of gynecology, prompted a change in the way involuntary childlessness was viewed, both popularly and medically. What had been regarded as barrenness, a personal misfortune, became infertility, a treatable condition. As Margaret Marsh and Wanda Ronner (1996) noted in their history of infertility in the United States, during the nineteenth century, involuntarily childless married women began to increasingly turn to medical expertise; by doing so, both clinicians and women accepted infertility as treatable. Infertility was now a recognized medical condition, but patients seeking treatment for it were still nearly all female (Marsh and Ronner, 1996).

Beginning in the late nineteenth century, Marsh and Ronner argue, women willingly underwent often invasive treatments, including various surgeries, to correct an impairment of their bodies in order to provide them with a chance for motherhood. By doing so, they sought a surgical restoration not just of their reproductive functions but also of their feminine identity. The convergence of motherhood as defining femininity with the increasing reliance on physicians to treat involuntarily childless women enabled a perception that infertility was a problem of and with the female body. This was reinforced as only women sought medical attention for infertility. For the majority of the nineteenth century, it was women alone who were considered, culturally and medically, to be infertile; men were only considered infertile if they were impotent (Marsh and Ronner, 1996). Infertility was characterized as a disease of the female body, and it was a disease that impaired a woman's ability to achieve her primary social role: motherhood (Apple, 1997; Marsh and Ronner, 1996). Women internalized infertility: a woman's inability to biologically bear children challenged the prevailing cultural norm that motherhood was the natural role for a feminine woman. Infertility, then, affected the way women saw their feminine selves.

As Marsh and Ronner (1996) found when they examined infertility in the late nineteenth and early twentieth centuries in the United States, doctors treated women across racial, ethnic, and class lines, suggesting the link between reproduction and femininity extended across the demographic spectrum. As the twentieth century wore on, increasing numbers of women sought medical expertise to enable them to conceive, with a sharp rise in infertility treatments accompanying the baby boom following World War II. During this intense pronatalist period in American history, to not be pregnant or have children within two years of marriage marked some women as odd – possibly even suggesting a lack of femininity. This tie between a woman's ability to conceive and her femininity, though ebbing a bit in the 1970s, remained strong through the course of the twentieth century (Marsh and Ronner, 1996; May, 1995).

5. Current cultural narratives: Women and infertility

The cultural connection between a woman's reproductive ability and her feminine identity has strong historical foundations in how women have been treated for infertility, foundations that continue to frame cultural conceptions of femininity as well as individual women's conception of themselves as female. Women have, and continue to, internalize their infertility. The contemporary literature on women without a cancer diagnosis but with

impaired fertility reveals a significant amount of stress and depression due to their condition. Women who are infertile but otherwise healthy are twice as likely to be depressed as fertile healthy women; indeed, these women report levels of psychological distress comparable to women with life-threatening illnesses (Davis and Dearman, 1991; Domar, Zuttermeister, 1993; Luske and Vacc, 1993). In her exploration of women's reactions to learning of their infertility, Gayle Letherby (2002) found that women experienced a profound shock to their sense of themselves, resulting in a challenge to their conception of identity as female.

What happens, then, to a woman's conception of herself when the option for biological motherhood is impaired or taken away because of her cancer diagnosis or cancer treatment? The existing literature on women whose fertility is impaired due to cancer treatment reveals similar psychological stress; for these women, "psychological distress may result from, not only the loss of the physical ability to conceive, but also a symbolic loss of the option or idea of fertility, regardless of whether this would have been acted upon or achievable" (Carter, et al., 2005, p. 93). The American Society of Clinical Oncology (ASCO) found that "surveys of cancer survivors have identified an increased risk of emotional distress on those who become infertile because of their treatment" (Lee et al., 2006, p. 2921). One study on young women with breast cancer found that "fertility concerns may complicate" their "treatment decision-making process" and that there is evidence these young women "may experience greater psychological distress and more difficulty with adjustment to the diagnosis and treatment of breast cancer" (Partridge, 2004, p. 4175). This actual or symbolic loss has potentially great implications for a woman's perceptions of herself as female, for motherhood is a culturally significant role most women see themselves in and which cancer potentially interrupts. Women whose cancer treatments threaten their fertility still want the experience of motherhood, most often biological motherhood (Lee et al., 2006, p. 2921). This equation of femininity with reproductive ability remains culturally resonant. Women today, as in the past, quite often tie their feminine identities to their reproductive capabilities; a cancer diagnosis is a recent, and additionally culturally powerful, component of a longer story concerning women, infertility, and medical treatment.

6. Historical foundations: Men and infertility

Women, as was just explored, were much more likely in both the nineteenth and the twentieth centuries to seek medical treatment for infertility. While this gender discrepancy was at first an outgrowth of the idea that only women were barren, and then that only women were infertile, it persisted even as physicians noted that there could be a male factor involved in infertility. Indeed, by the late nineteenth century, some gynecologists began calling for husbands to be examined when wives came seeking a cure for their childlessness. But even as the concept of barrenness gave way to infertility for women in the nineteenth century, very few doctors asked to examine a patient's husband, as most physicians still considered it rare for a virile man to be infertile (Marsh and Ronner, 1996). This cultural conception was reflected and reinforced medically in the nineteenth century. In the early twentieth century, after examining a young woman who had been married for six years and came to him for "relief" of her suffering from infertility, physician Gustavus Blech (1903) took a swab of semen from the young woman's vagina and looked at it under a microscope. He found "not a spermatozoon in sight." But though he noted this physiological matter, and

recommended to other physicians the need to look at the sperm of husbands' when a woman came to them for the treatment of infertility, he also stressed the importance of whether or not the man was "capable of performing the sexual act" when analyzing if the infertility was male or female factor (Blech, 1903, 45-46).

As the Blech case illustrates, male fertility was connected to virility through the twentieth century, even though physicians acknowledged that the viability of the semen was an identifiable factor. Though by the 1930s physicians recognized that infertility was comparable in both genders, and by the 1950s semen analysis had become a standard protocol for infertility intervention, women largely remained the focus of medical fertility treatment. More than one physician lamented in the first half of the twentieth century that too many of his peers were eager to first perform invasive procedures (such as surgery) on the wife before testing the viability of the husband's sperm (Marsh and Ronner, 1996; May 1995). But even with the knowledge that men were infertile as often as women, this gender difference in treatment continued for the course of the twentieth century. In 1963, the president of the American Society for the Study of Sterility (later called ASRM) explained at the group's annual meeting that such a protocol remained justified because many infertility specialists believed men were rarely infertile. But just as important, since women usually initiated medical treatment, their husbands were often left unexamined (May, 1995). So while the physiological knowledge was there and the understanding that men were as likely to be infertile as women was available, women were the ones being treated because they sought the medical intervention. Even if a physician could not find a physiological reason on the woman's body, both the woman and the physician frequently believed the impairment was with her body. From the nineteenth century to the present, women have most commonly sought medical intervention for infertility, often doing so in order to spare their husbands from possible humiliation – a humiliation based on the idea that an infertile man was an impotent man (Marsh and Ronner, 1996; May, 1995).

While femininity was tied to reproduction, and this tie held both culturally and medically over the course of the twentieth century, masculinity, in contrast, was tied to sexual virility. Here we are extending the work of Marsh and Ronner (1996) in their history of infertility in the United States to argue that there is a historical basis for the current differences of fertility-related distress among male cancer patients compared to female cancer patients: men's medical treatment was not so tied, and did not reinforce, infertility as an impairment of their bodies. When infertility was associated with men, it was generally seen as an impairment with virility.

7. Current cultural narratives: Men and infertility

Virility remains a significant component of cultural conceptions of masculinity and men's gendered identities. Sexual prowess is often seen as a way of proving one's masculinity. Furthermore, the male genitals are generally central to a man's coherent sexual identity (Gurevich et al., 2004), and are associated with stereotypical masculine traits like "strength" and "courage" (Szasz, 1998). Because of the personal, as well as social, significance of the male genitals, having "misfunctioning" (e.g. impotent, prematurely ejaculating, infertile) genitals or genitals that look "abnormal" (e.g. small penis, missing a testicle) can diminish men's sense of masculinity. For instance, in a ranking of the most humiliating experiences for college age men, the top three had to do with sexual function and the appearance of

genitals: 1. unable to maintain an erection during sex; 2. losing a testicle to cancer; 3. being teased about penis size (Mormon, 2000).

This trend holds true for men facing cancer, especially those with cancer affecting the genitals: listing the second most humiliating experience as losing a testicle to cancer highlights the deleterious effects cancer can have on men's self-worth and identity. A qualitative study on men with testicular cancer found that definitions of masculinity continue to be strongly tied to sexual performance and the appearance of "normal" genitals, both of which can be threatened by cancer and cancer treatments (Gurevich et al., 2004).

Whereas the centrality of virility and male genitals to men's sexual identity is supported in the literature (Gurevich, et al., 2004), the importance of fertility for men—the desire to have biological children and the role fertility plays in their identity—is not just ambiguous, but is often contradictory. As previously discussed, some studies found male cancer patients value their fertility as much as female cancer patients whereas other studies show that women value their fertility much more than men. Much of the broader literature on infertility in the general population, not just cancer patients, supports the latter finding. In his interviews with mainly white, middle-class, heterosexual married couples facing infertility,[1] Arthur Greil (1991) found that the husbands were more likely to view the experience of infertility as disappointing, though not as a threat to their identity. The wives, in contrast, saw infertility as devastating, something that spoiled their identities and signified their role failure as woman, wife, and mother (Greil, 1991). Furthermore, in a review of the literature, Greil discusses how both qualitative and quantitative studies show that women react more negatively to infertility (e.g. have lower self-esteem, blame themselves for their infertility, feel defective, etc.) than men (Greil, 2010). Even when the couple is suffering from male factor infertility, most of the literature concludes that this does not seem to change men's response to infertility (Greil, 1997).

Given that infertility is not as devastating for men as impotence or abnormal genitals, it is not surprising that a diagnosis of sterility ranks lower (in fifth place) on the list of the most humiliating experiences for college men than experiences that are more closely connected to sexual performance and the appearance of "normal" genitals (Mormon, 2000).[2] It is worth noting that men older than traditional college age students (18-22 years old) may have come up with a different ranking; while college age men are typically trying to avoid fatherhood and consequently do not value fertility as much at this stage in their lives, older men are probably more interested in becoming a father and thus may find a diagnosis of sterility more troubling.

While sterility and infertility may not top the list as the most humiliating experience for young men, they are still negative experiences. A large reason for this is the close relationship between men's virility and fertility (Inhorn, 2002). The historical association of infertility with impotence and the importance of virility to current conceptions of

[1] White, middle-class, heterosexual married couples are the most common participants of many empirical studies on infertility in part because they are the social group most likely to seek out infertility treatment. In the last handful of years, however, more research has been done on people of color and infertility. See, for example, Becker et al., Jain, and White et al.

[2] Having a rectal exam ranks fourth. While this experience does not directly deal with male sexuality, being penetrated is usually associated with women, which is in part why this experience is humiliating.

masculinity result in cultural understandings of infertility as an indicator of emasculation. Cynthia Daniels (2006), a political scientist researching the politics of reproduction, asserts that "the ability to biologically father one's children remains a hallmark of one's manhood, and infertility remains a source of masculine shame" (p.161). Specifically, questions about the viability of sperm production incite questions about the viability of masculinity (Gurevich 2004; Daniels 2006) This association is reflected in the secrecy and stigma surrounding heterosexual couples using sperm donation and results in practices like "matching" sperm donors to the physical traits and characteristics of the social father to hide his infertility (Becker, 1994; Daniels, 2006).

In addition to threatening masculinity, infertility may adversely affect men because it leaves their desire for biological children unfulfilled. Yet, this reason may play a smaller role for men than for women. The cultural pressure for women to have biological children and the fact that motherhood is an important part of many women's identity are thought to be a significant factor in why infertility is so devastating to women. That "[t]here is, in American society, no 'fatherhood mandate' with the same force and intensity as the 'motherhood mandate'" may explain, at least in part, men's less strong reaction to infertility (Greil 1991, p. 64).

While there is still less pressure for men to be biological fathers compared to the pressure women feel to be mothers (at least in the U.S.), men are much more active in their children's lives than even a generation ago. We see this change reflected in new concepts and policies. For example, the concept of a stay-at-home dad is relatively recent. This term probably did not exist, or at the very least was not ubiquitous, a couple of generations ago. Today, the U.S. Census Bureau estimates that there are 143,000 stay-at-home dads. Men's involvement in primary caretaking of children is also seen in policies like extending maternity leave to men through paternity leave. A recent study shows that 89% of men took some time off after the birth of their child (Nepomnyaschy and Waldfogel, 2007).

Men's increased involvement in their children's lives shows not only that fatherhood is extremely significant for many men, but also that men's active participation in their children's lives is becoming more socially acceptable. Indeed, gender norms surrounding fatherhood are changing, which influences the value men place on becoming biological fathers. Fatherhood today seems to play a greater role in men's identity and their vision of an ideal life. This social change helps explain the mixed results on the importance of fertility to male cancer patients as illustrated in the quality of life surveys. We are in a time of transition in which more men are value active fatherhood, but there are still men who adhere to traditional gender roles. The survey results mirror this transition, as some surveys show that men are less interested in parenthood than women and other studies reveal that men's and women's interest in fertility are equivalent. If men's interest in fatherhood continues to grow, we can expect that in the future, studies will uniformly find that women and men equally value parenthood.

8. Clinical implications and conclusions

The persistent gender bias – women's bodies demand medical intervention while men's bodies are left alone as long as there is a physical sign of sexual potency – is a historically important cultural phenomenon that continues to shape current conceptions of femininity, masculinity, and fertility. Understanding the history of the differences in medical treatment

of infertility provides a unique contextualization of contemporary cancer patients' views on their fertility and the possibility of their future infertility. Quality of life studies consistently show that fertility is very important to female cancer patients and while the results for male cancer patients are mixed, many men do strongly value their fertility. These social science studies, however, are often not translated into clinical practice by health care providers. Instead, many providers continue to make assumptions "based on the patient's age, sex, diagnosis, culture, and partnership status without checking with the patient" (Horden and Street, 2007, p. 227). Furthermore, providers' personal characteristics (e.g. age, sex, etc.) can also influence whether they discuss fertility preservation treatment with their patients. For example, a recent study found that female oncologists were more likely to refer their patients to a reproductive endocrinologist or infertility specialist compared to male oncologists. While discussions and referrals for fertility preservation among adult cancer patients are improving, they remain suboptimal: only forty-seven percent of respondents always or often refer cancer patients of childbearing age to a reproductive endocrinologist or infertility specialist (Quinn, 2009). Health care providers need to openly address potential infertility as a consequence of cancer treatment so that their patients are informed about and offered fertility preservation options.

By contextualizing female and male cancer patients' views on infertility within a historical framework, we have shown that these differences in views are not innate, but rather are shaped by gender norms, both historical and current. Understanding the social factors that influence people's views on infertility will enable health care providers to better aid their patients in their fertility preservation decision making. In other words, situating patients within their social environment, rather than seeing them as free floating individuals or reducing them to their diseases, allows providers to acknowledge the various social factors that contribute to and are at stake in decisions about fertility preservation treatment. Familiarity with gender norms as well as recognition they are changing (especially the role of fatherhood for men) equips providers with the ability to understand the discrepancy between women's and men's views and the discrepancy among men's views on infertility. This can help providers be more empathetic to their patients' needs and concerns, while at the same time not pigeonholing patients by sex/gender.

Various health organizations, including ASCO and ASRM, have issued guidelines on fertility preservation to their practitioners for all cancer patients stating that "the available evidence suggests that fertility preservation is of great importance to many people diagnosed with cancer" (Lee, 2006, p. 2921; Ethics Committee of ASRM, 2005). While it is important for these organizations to set guidelines, the mere existence and dissemination of guidelines may not be enough to make substantial changes in provider referrals for fertility preservation treatment. A deeper appreciation for the social factors involved in patients' views on infertility may engender greater change and specifically more referrals for fertility preservation treatment. Safeguarding fertility for cancer patients should be considered as important of a quality of life issue as breast reconstruction and hair replacement (Campo-Engelstein, 2010), and the options for fertility preservation should be part of the discussion regarding a patient's cancer treatment – regardless of gender. Doing so respects both women's and men's desire to be – or at least retain the possibility of being – biological parents.

The best way to ensure that these historical foundations and social factors are integrated into clinical care is through the development of an appropriate intervention. By creating

both provider and patient education materials that contextualize fertility preservation choices within the social environments in which they occur, as well as offering patients coping strategies that recognize their unique gendered responses, we can help address relevant patient concerns and enhance decision-making capabilities. Jordan et al. (1999) has recognized that significant gender differences exist in coping strategies for infertility, and these differences need to be addressed in the development of patient education materials that offer coping strategies (Jordan, 1999). Both historically and presently, infertility has been described as a devastating diagnosis for cancer patients, both female and male. Recognizing and outlining the historical influences that have contributed in creating this distress, as well as the current social factors that reinforce it, is a step in the right direction. Incorporating these analyses into the development of interdisciplinary patient and provider education materials is the next step towards translating these findings into clinical practice.

9. Acknowledgement

This research was supported by the Oncofertility Consortium NIH 8UL1DE019587, 5RL1HD058296.

10. References

American Cancer Society. (2006). Cancer facts and figures 2006. Atlanta: American Cancer Society.

American Cancer Society. (2008). Cancer facts and figures 2008. Atlanta: American Cancer Society.

American Cancer Society. (2009). Cancer statistics 2009 presentation. http://www.cancer.org/docroot/PRO/content/PRO_1_1_Cancer_Statistics_2009_Presentation.asp (October 12, 2009).

Apple, R. D., Golden, J. (1997). Introduction. In R. D. Apple and J. Golden (Eds)., Mothers and motherhood: readings in American history. Columbus: Ohio State University Press.

Bardwell, W. A., Profant, J., Casden, D. R., Dimsdale, J.E., Ancoli-Israel S., Natarajan, L., Rock, C.L., Pierce, J.P., and Women's Healthy Eating & Living (WHEL) Study Group. (2008). The relative importance of specific risk factors for insomnia in women treated for early-stage breast cancer. *Psychooncology*, *17*(1), 9-18.

Becker, G., M. Castrillo, Rebecca Jackson et al. (2006). Infertility among low-income Latinos. *Fertility & Sterility,* 85(4): p. 882-7.

Becker, G. and R. Nachtigall. (1994) 'Born to a mother': The cultural construction of risk in infertility treatment in the U.S. *Social Science & Medicine, 39*(4), 507-518.

Blech, G. M. (1903). *The Practitioner's guide to the diagnosis and treatment of diseases of women.* Chicago: M. Robertson.

Carter J, Rowland K, Chi D, et al. (2005). Gynecologic cancer treatment and the impact of cancer-related infertility. *Gynecol Oncol, 97*, 90-95.

Campo-Engelstein, L. "Consistency in Insurance Coverage for Iatrogenic Conditions Resulting from Cancer Treatment Including Fertility Preservation." Journal of Clinical Oncology 28.8 (March 10, 2010)

Carver, C. S. (1997). You want to measure coping but your protocol's too long: Consider the Brief COPE. *International Journal of Behavioral Medicine*, *4*, 92-100.

Chang V.T., Hwang S.S., Feuerman M., Kasimis B.S., Thaler H.T. (2000). The memorial symptom assessment scale short form (MSAS-SF). *Cancer, 89*(5), 1162-71.

Clayman, M. L., Galvin, K. M., and Arnston, P. (2007). Shared decision making: Fertility and pediatric cancers. In Woodruff, T. K. and Snyder, K. A., eds. *Oncofertility: Fertility preservation for cancer survivors*. New York: Springer.

Connell S., Patterson C., Newman B. (2006). A qualitative analysis of reproductive issues raised by young Australian women with breast cancer. *Health Care Women, 27*, 94-110.

Crawshaw, M. A., Sloper, P. (2010) 'Swimming against the tide' – the influence of fertility matters on the transition to adulthood or survivorship following adolescent cancer. *European Journal of Cancer Care*, Jan 19.

Daniels, C. R. (2006). *Exposing men: The science and politics of male reproduction*. New York, Oxford University Press.

Davis, D. C., Dearman, C. N. (1991). Coping strategies of infertile women. *Journal of Obstetric, Gynecologic, & Neonatal Nursing, 20* (3), 221-8.

DevCan: Probability of Developing or Dying of Cancer Software, Version 6.3.0 Statistical Research and Applications Branch, NCI, 2008. http://srab.cancer.gov/devcan.

Duman, A. D., Zuttermeister, P.C., Friedman, R. (1993). The psychological impact of infertility. *Journal of Psychosomatic Obstetrics and Gynecology*, 14 (Suppl S. Dec.), 45-52.

Dunn J., Steginga, S. K. (2009). Young women's experience of breast cancer: Defining young and identifying concerns. *Psychooncology*, *9*, 137-146.

Ethics Committee of the American Society for Reproductive Medicine. (2005). Fertility preservation and reproduction in cancer patients. *Fertility and Sterility, 83* (6), 1622-1628.

Green D. H., Galvin, H., and Horne, B. (2003). The psycho-social impact of fertility on young male cancer survivors: A qualitative investigation. *Psychooncology, 12*,141-152.

Greil, A. L. (1991). *Not yet pregnant: Infertility couples in contemporary America*. New Brunswick: Rutgers University Press.

Greil, A. L., et al. 2010. The experience of infertility: A review of recent literature. *Sociology of Health & Illness, 32*(1):140–62.

Gurevich, M., Bishop, S., Bower, J., Malka, M., and Nyhof-Young, J. (2004). (Dis)embodying gender and sexuality in testicular cancer. *Social Science & Medicine*, *58*, 1597-1607.

Heinemann, K., Ruebig, A., Potthoff, P., Schneider, H., Strelow, F., Heinemann, L. and Minh Thai, D. (2004). The menopause rating scale (MRS) scale: A methodological review. *Health and Quality of Life Outcomes*, *2*, 45-52.

Horden, A. J., and Street, A. F. (2007). Communicating about patient sexuality and intimacy after cancer: Mismatched expectations and unmet needs. *Medical Journal of Australia, 186*(5), 224-227.

Horner, M. J., Ries, L. A. G., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S. F., Feuer, E. J., Huang L., Mariotto, A., Miller, B. A., Lewis, D. R., Eisner, M. P., Stinchcomb, D. G., Edwards, B. K. (eds). SEER Cancer Statistics Review, 1975-2006, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2006/, based on November 2008 SEER data submission, posted to the SEER web site, 2009.

Jain, T. (2006). Socioeconomic and racial disparities among infertility patients seeking care. *Fertility & Sterility,* 85(4): p. 876-81.

Jeruss, J. S. and Woodruff, T. K. (2009). Preservation of fertility in patients with cancer. *The New England Journal of Medicine*, *360*(9), 902-911.

Jordan, A., and Revenson, T. A. (1999). Gender differences in coping with infertility: A meta-analysis. *Journal of Behavioral Medicine*, *22*(4), 341- 358.

Kinahan, K. E., Didwania, A., and Nieman, C. L. (2007). Childhood cancer: Fertility and psychosocial implications. In Woodruff, T. K. and Snyder, K. A. (eds). *Oncofertility: Fertility preservation for cancer survivors*. New York: Springer.

Letherby, G. (2002). Challenging dominant discourses: Identity and change and the experience of 'infertility' and 'involuntary childlessness'. *Journal of Gender Studies, 11*, 277-288 (ref. on pg. 279).

Lee, S. J., Schover, L. R., Partridge, A. H., Patrizio, P., Wallace, W. H., Hagerty, K., Beck, L. N., Brennan, L. V., Oktay, K. (2006). American Society of Clinical Oncology recommendations on fertility preservation in cancer patients," *Journal of Clinical Oncology, 24* (18), 2917-2931.

Loscalzo MJ, Clark KL. The psychosocial context of cancer-related infertility. In: Woodruff TK, Snyder KA, editors. Oncofertility: Fertility Preservation for Cancer Survivors. New York: Springer, 2007:180-190.

Lukse, M. D., and Vacc, N. A. (1999). Grief, depression, and coping in women undergoing infertility treatment. *Obstetrics & Gynecology, 93* (2), 245-51.

May, E. T. (1995). *Barren in the promised land: Childless Americans and the pursuit of happiness.* New York: Basic Books.

Mormon, M.T. (2000). The influence of fear appeals, message design, and masculinity on men's motivation to perform the testicular self-exam. *Journal of Applied Communication Research 28*, 81-116.

National Cancer Institute. (2006). Facing forward: life after cancer treatment. Washington, DC: U.S. Department of Health and Human Services, National Institutes of Health.

Nepomnyaschy, L. and Waldfogel, J. (2007). Paternity leave and fathers' involvement with their young children: Evidence from the American ECLS-B. *Community, Work & Family*, *10*, (4), 427 – 453.

Patridge, A. H., Gelber, S., Peppercorn, J., Sampson, E., Knudsen, K., Laufer, M., Rosenberg, R., Przypyszny, M., Rein, A., Winer, E. P. (2004). Web-based survey of fertility issues in young women with breast cancer. *Journal of Clinical Oncology*, *22*, 4174-4183.

Quinn, G. P., Cadaparampil, S. T., Lee, J. H., Jacobsen, P.B., Bepler, G., Lancaster, J., Keefe, D. L., and Albrecht, T. L. (2009). Physician referral for fertility preservation in oncology patients: A national study of practice behaviors. *Journal of Clinical Oncology*, *27*(35), 5952-5967.

Radloff, L.S. (1977). The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Rieker, P. P., Fitzgerald, E. M., Kalish, L. A. (1990). Adaptive behavioral responses to potential infertility among survivors of testis cancer. *Journal of Clinical Oncology*, *8*, 347 – 55.

Saito, K., Suzuki, K., Iwasaki, A., Yumura, Y., Kubota, Y. (2005). Sperm cryopreservation before cancer chemotherapy helps in the emotional battle against cancer. *Cancer*, *104*,521-524.

Schover, L. R. (1999). Psychosocial aspects of infertility and decisions about reproduction in young cancer survivors: a review. *Medical and Pediatric Oncology*, *33*:53-59.

Schover, L.R., Rubicki, L.A., Martin, B.A., Bringelsen, K.A. (1999). Having children after cancer. A pilot survey of survivors' attitudes and experiences. *Cancer, 86*, 697-709.

Schover, L.R., Brey, K., Lichtin, A., Lipshultz, L.I., Jeha, S. (2002a). Knowledge and experience regarding cancer, infertility, and sperm banking in younger male survivors. *Journal of Clinical Oncology, 20*, 1880-1889.

Schover, L.R., Brey, K., Lichtin, A., Lipshultz, L. I., Jeha, S. (2002b). Oncologists' attitudes and practices regarding banking sperm before cancer treatment. *Journal of Clinical Oncology, 2*, 1890-1897.

Szasz, I. (1998). Masculine identity and the meanings of sexuality: A review of research in Mexico. *Reproductive Health Matters, 6*(12), 97-104.

Tschudin, S., Bitzer, J. (2009). Psychological aspects of fertility preservation in men and women affected by cancer and other life-threatening diseases. *Human Reproduction Update, 1:* 1-11.

U.S. Census Bureau. Press release: Father's day. http://www.census.gov/Press-Release/www/releases/archives/facts_for_features_special_editions/006794.html. Published June 12, 2006. Accessed March 11, 2010.

Ware, J. E., Sherbourne, C. (1992). The MOS 36-Item short-form health survey (SF-36): A conceptual framework and item selection. *Medical Care*, *30* (6), 473-483.

Wasserman, A. L., Thompson, E. I., Wilimas, J.A., and Fairclough, D.L. (1987). The psychological status of survivors of childhood/adolescent Hodgkin's Disease. *American Journal of Diseases of Children, 141*:626-631.

Wenzel, L., Dogan-Ates, A., Habbal, R., Berkowitz, R., Goldstein, D.P., Bernstein, M., Khusman, B.C., Osann, K., Newlands, E., Secki, M.J., Hancock, B., Cella, D. (2005). Defining and measuring reproductive concerns of female cancer survivors. *Journal of the National Cancer Institute Monographs,* 34, 94-98.

White, L., J. McQuillan, and A.L. Greil. (2006). Explaining disparities in treatment seeking: The case of infertility. *Fertility & Sterility*, 85(4): p. 853-7.

Zanagnolo, V., Sartori, R., Trussardi, E., Pasinetti, B., Maggino, T. (2005). Preservation of ovarian function, reproductive ability and emotional attitudes in parents with malignant ovarian tumors. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 12*3, 235-243.

Zebrack, B. J., Casillas, J., Nohr, L., Adams, H., Zelter, L. K. (2004). Fertility issues for young adult survivors of childhood cancer. *Psychooncology, 13*, 689-699.

Zelter, L. K. (1993). Cancer in adolescents and young adults: Psychosocial aspects in long-term survivors. *Cancer, 71*(Suppl 10), 3463-3468.

Edited by Jay Maddock

Public health can be thought of as a series of complex systems. Many things that individual living in high income countries take for granted like the control of infectious disease, clean, potable water, low infant mortality rates require a high functioning systems comprised of numerous actors, locations and interactions to work. Many people only notice public health when that system fails. This book explores several systems in public health including aspects of the food system, health care system and emerging issues including waste minimization in nanosilver. Several chapters address global health concerns including non-communicable disease prevention, poverty and health-longevity medicine. The book also presents several novel methodologies for better modeling and assessment of essential public health issues.

ISBN 978-953-51-7007-5

Photo by malija / iStock

IntechOpen