Reviewers´ statements:

The publication is important for improvement of knowledge and orientation in the statistical analyses for specialists in healthcare and medicine sciences and for further research in this area. The publishing is very current, considering the absence of such complex monograph dedicated to the statistics used in health and healthcare policy in Slovakia. The scientific monograph presents high-quality elaborated subject matter of statistical procedures applied to the evaluation and analysis of data and information in the area of health condition, disease and death rate of the population groups, as well as utilization of economic indicators and mathematical models in the public healthcare system.

Prof. MUDr. Ľudmila Ševčíková, CSc., Institute of Hygiene, Faculty of Medicine, the Komensky University in Bratislava

The authors present a complex view on the statistics of health and healthcare, comprehensively and thoroughly elaborated and enriched with the latest findings. The publication presents methods and major approaches of data processing that enable interpretation of results and their application to the development of public healthcare strategy. I recommend the book as a good information source about specific subject matter – health and healthcare.

prof. MUDr. Henrieta Hudečková, PhD, MPH, head of the Public Healthcare Institute, Jessenius Faculty of Medicine, the Komensky University in Martin

The publication meets all criterions of scientific monograph. In a friendly way based on particular examples and pertinent illustrations / pictures, it explains demanding and often-neglected subject matter that is very important especially in the present times for the understanding of huge amount of statistical data in majority of medicine specializations and healthcare management.

prof. MUDr. Peter Krištúfek, CSc., The Pneumology and Phtiseology Clinic of the Faculty of Medicine, the Slovak University of Healthcare

ISBN 978-80-8082-732-8

MarekPsota

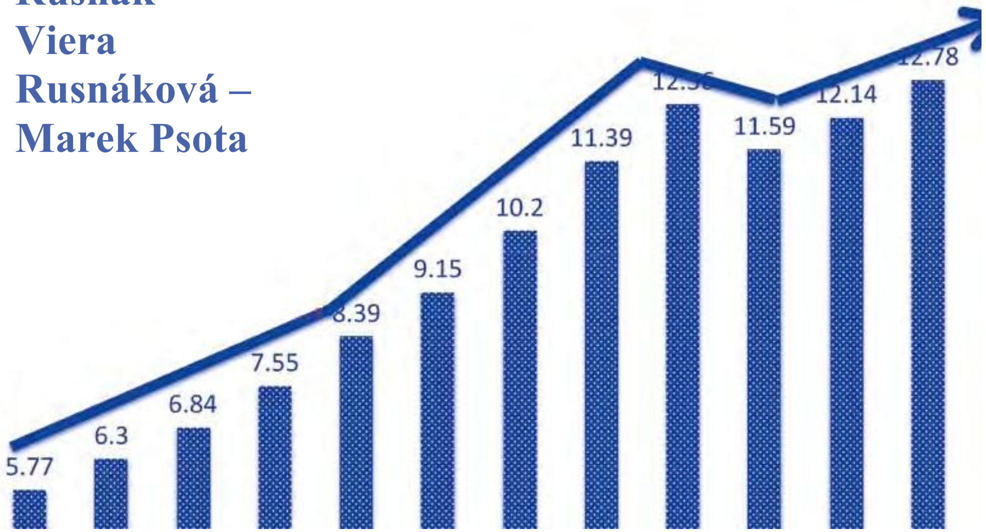
Statistics of Health

Viera Rusnáková MartinRusnák

9 788080 827328 >

Statistics of Health

Martin Rusnák – Viera Rusnáková – Marek Psota



TRNAVA 2013

Faculty of Healthcare and Social Work

If you cannot measure it, you cannot improve it.

Lord Kelvin

Reviewers:

prof. MUDr. Ľudmila Ševčíková, PhD.
prof. MUDr. Henrieta Hudečková, PhD.
prof. MUDr. Peter Krištúfek, CSc.

Statistics of Health

Martin Rusnák – Viera Rusnáková – Marek Psotka

TRNAVA 2013

ISBN 978-80-8082-732-8

Table of Contents

Chapter 1

Introduction: From Data to Information

Publication objectives

Many areas intersect in the statistics of health, namely methods of the population study in the area of health and disease – demography, epidemiology, healthcare and economic statistics. This includes also health and healthcare, and health policy as a tool of synthesis of views obtained from the quantitative studies, combining the studies´ outcomes with political, ethical and personal values of people creating the policy.

Healthcare and health statistics also mirror or provide feedback for evaluation of the effects of changes demonstrated in the area of public health as a result of targeted intervention. Compiling the publication, we referred mainly to the experiences with the public healthcare and management students´ preparation. We were aware of missing book on the Slovak market that would help orientate oneself in the area of health statistics. While this topic wasn´t of a key importance in Slovakia during the previous political regime, its position has significantly improved after the changes occurred in the healthcare services organization and management. The subject matter was sufficiently described in a chapter of the book Social Medicine [1] in the past but currently it should be updated and completed. Foreign literature has paid much attention to the health statistics; however we suppose that such publication should be issued also in Slovakia, taking in account continuous limitations resulting from insufficient knowledge of foreign languages. We prepared the publication including specification of foreign literature sources, hoping that we will provide a reader with ideas and grounds for further study. We didn´t aim at providing a complex guide in the area of statistics, an encyclopedic but rather strived for outlining trends, presentation of the basis and examples of processing and interpretation.

The publication also aims at presenting the data processing methods with focus laid on approaches enabling interpretation of the processing results, and thus providing the information. Considering a complex nature of the data rather easily available in the present times thanks to the Internet, the authors made efforts to present the methods and interpretation of results so as they can be applied to the study and development of strategies in the area of public health.

From data to information

There is kind of admiration of the events quantification in the health sciences. Disease is often defined as condition when certain parameters are out of regular range. For example, systolic blood pressure values are higher than 140 torr. It is not easy to specify the limit value deciding on whether we say a man that he is ill or calm him down, saying that the parameter is within acceptable range (hereinafter "he, him, his" shall refer to both genders). We will talk about it in detail in the chapter dealing with diseases. Old age is also defined as certain age passed through, etc. Hospitals, public healthcare offices, agencies dealing with the statistics release the tables full of figures. The meaning of such efforts is often lost and could seem useless for a not involved outsider. Years ago, many of us were fascinated with the book[1] about information, written by neurologist William Ross Ashby. [2] The book explains principles of informatics on the life examples (e.g. how the Indians forward messages one to another via smoke from flame). At the time, the informatics was called cybernetics. The book is very reader-friendly and, even if it could seem outdated today, we recommend it to everybody who is seriously interested in the informatics. The book is a follow up of fundamental work of Norbert Wiener. [3] These books weren´t ready available in the past since the cybernetics was rejected by the Soviet philosophers in the 50s of the 20th century and thus successfully contributed to the underdevelopment of countries they had control over, (thus also Slovakia) compared to the rest of the world.

Data

Let´s start with data. Data refer to any report regardless having any information content for us or no; in other words, whether it says anything new to us or no. Data refer to reports expressing some facts about processes or elements in the real world.

---

[1] Available freely on http://pcp.vub.ac.be/books/IntroCyb.pdf

Data refer to letters, numbers, words, signs or the combination of all.

Information

To understand why the mankind tries to express certain events and situations with a number, we have to explain the difference between data and information. Information is a report on occurred one of possible events amongst the variety of existing effects, resulting in reduced lack of knowledge about the effect at the information recipient. It is an idea presented in particular language (through symbols), expressing certain object status and behavior. Information can be measured, i.e. expressed through numbers. Nowadays we commonly express the information quantity with bites, bytes and their multiplications. Along with quantity, information is characterized by recipient to whom it can but doesn´t have to have any meaning.

Thus, we can say that it isn´t enough to only have data like health and illness statistics. The data itself are not useful enough and become such after interpretation and processing, then called *information*. Accordingly, all information must be concurrently data but not every data represents interpreted facts. Speaking of data, these are facts collected by people based on their observations and experiences. They are raw facts. Majority of data refers to simple records of facts or observations. They are recorded on paper of stored in memory, e.g. birth dates of our relatives. Information represents what the data convert to after processing and interpretation. We can also say that data are based on observation while information is based on research. Information is of a more material nature. It is more reliable, supported by research and generated by experts, scientists and researchers. Anyway, we would like to make you aware of the phenomenon called *datism*. This expression presents lack of common sense describing collecting and request for data not required for fulfillment of goal, regardless the goal.

Indicator

Term "indicator" has been used in professional terminology and common language; while we don´t always recognize its actual meaning. There are more than one definitions of "indicator". In general, we can say that indicator provides us with a proof of evidence that certain term exists or certain results have or haven´t been reached. Properly chosen indicator helps understand where we are, where we go to and how far we are from the target. Indicator can refer to a sign, number, or graphical expression. Indicators usually express results of measurement. They are pieces of information summing up the properties of systems, or describing what´s going on in the system. As such, indicators serve to make decision or assess progress towards set forth goals, outputs, or results. They are usually aggregated raw data that can be further grouped together and create complex indicators. For example, consider health data, namely headcount of the dead. If we analyze them, they will become health information. [4] If the information was interpreted and conclusions were drawn, we will obtain assessment – health intelligence [2]. It is important for deduction of informed decisions. Thus, indicators represent a single segment in the chain of outputs used during health intelligence or health services improvement. Typically, indicators consist of the combination of two and more data; for example number of deaths per citizens´ headcount or Body Mass Index (weight, height and age). Often it is stated that an indicator is a ratio that is true but it doesn´t have to be proportion of two figures only. Demography used to distinguish analytic data as extensive proportional figures, called indicators. Furthermore, intensity data are distinguished, so called measures or quotients. And finally, there are proportional comparable figures, called indexes. [5] For purposes of this publication, we will use the term "indicator". Complex indicators usually consist of three and more figures. Proper indicator should meet a few requirements (see the Table No. 1):

1. Reliability/ reproducibility - repeated measurements at similar circumstances, performed by the same or various persons should produce similar results..
2. Validity - indicator measures such properties or characteristics that the measurement was aimed at.
3. Sensitivity - differences or changes can be identified on fine enough level required by the users.
4. Acceptability - presumed users should consider the indicator understandable, credible and useful.
5. Feasibility – data can be collected without useless administration or financial burden.
6. Universality – flexibility: indicator can be adjusted to various populations or circumstances.

Table No. 1: Required properties of indicator, based on [6]

Particular indicators will be described in detail in the following chapters according to their primary purpose. Selection was made according to the combination of the indicator origin, i.e. on basis of the type of primary data that the indicator was determined to and which characteristics it describes. It is not always possible to avoid repetition whether decision on the category should be adapted to either complexity or understandability of the text.

2 The term "intelligence" should be understood as learning, measurement, and evaluation of health but the intellect only. With this term, we can describe the methods of obtaining and application of knowledge that supports decision-making related to citizens´ health improvement.

Specific aspects of health and healthcare statistics

Statistics of health and health-related services has become a natural part of quantitative expression and research of population health and health-related factors. Traditional approach to the public health deals with population health while the concept of new public health deals with health of individuals and groups of population [7]. This fact has extended the approach with bio-statistics and related areas. In this environment, we understand the *health statistics* as a tool for documentation of the population health condition and specific groups defined within the population. It helps identify the differences at health condition, approach to and utilization of healthcare services expressed according to ethnicity, social-economic status, geographical location, age, gender and other population characteristics. It is necessary for monitoring and recording of experiences of healthcare services providers and recipients, and monitoring of health and service trends. It is impossible to characterize the population health problems without healthcare statistics, which is also necessary for the measures, policies and programs effects evaluation. Last but not least, it is correlated to the biomedicine research and research of healthcare services.

Bio-statistics and Statistics of Health

It is not always distinguished between bio-statistics and health statistics in the healthcare environment on daily basis. Bio-statistics deals with application of statistical methods to biological problems solving; we used to distinguish also between medical statistics in the medical environment. On the contrary, *Statistics of Health* deals with aggregated data describing and enlisting features, events, behavior, services, sources, results or cost in relation to health, disease and healthcare services. The data can be obtained from the studies, patients´ records and administrative documents. *Demographical statistics* (monitoring of population life) represents its variable subset. [8] Simpler definition of health statistics states that it represents numerical data characterizing population health and effects affecting it. [9] Of course, the first definition, despite of being broader, provides more accurate specification and we will use it in this document. Nevertheless, we cannot completely avoid intersection with various areas co-participating on creation of a complex picture. Let´s include three areas herein: public health surveillance, public health informatics and population health statistics. Students of specialization "Epidemiology" used to use the term surveillance within classical terminology as continuous control of all aspects of disease development and spread, being subject to efficient control. [10], [11] Currently the term has been ever more used in the original meaning of the term surveillance – public health surveillance. It should be understood as continuous and systematic collection, analysis, interpretation and spread of data about the effects in relation to health, serving for application on the public healthcare system, for example measures aimed at reducing illness and death rate and improving health condition. Surveillance represents a part of at least eight functions of the public health, including support of disease identification and interventions, estimated consequences of disease or injury, support of illustration of natural health condition development, determination of disease distribution and spread, creation of hypotheses and encouraging of research; evaluation of preventive and control, measures and simplification of planning process. Identification of increased disease occurrence compared to regular status represents another important function of the public health surveillance. 12] Various scientific specializations contribute to such broad interpretation of surveillance, for example demography and epidemiology, healthcare systems research and clinical epidemiology.

Informatics

Informatics represents an independent section, especially informatics in the area of public health. American authors [13] define informatics as systematic use of informatics and computer sciences and technology in the area of public health practice, research and education. This scientific discipline outreach is defined as conceptualization, proposal, development, application, improvement, sustaining and evaluation of communication systems, surveillance systems, information and educational systems relevant for the public health. The definition clearly indicates that the two disciplines intersect mainly in the area of data collection and processing. Current databases (DB) contain all relevant data that we will use in this book. Many of them are available directly on the Internet. Program application process expects very good knowledge and skills in the area of informatics for purpose of data collecting and processing, and data processing that uses significantly complicated statistical methods. Last but not least, public evidence used in the practical public healthcare represents an outcome of information processing through high-specialized and sophisticated procedures. [14] Both scientific research and day-to-day practice expects and requires significant degree of skills when using them. All the mentioned reasons also imply the need for informatics training where the public health or related subjects are taught.

Demography

Demography as the science dealing with human population reproduction study represents the basis for population health knowledge. Demography covers all events and processes associated with human population reproduction. Demographical statistics deals with quantification of demographic events and processes [15].

Epidemiology

Epidemiology and mainly related epidemiological statistics represents also very important part thereof. We can find various definitions in the literature but one of them provides for main terms and concepts that characterize modern epidemiology today.

Epidemiology is the study of distribution and determining aspects of statuses or events related to health of certain population and application of the study to health problems solution.

Definition 1 Epidemiology pursuant to [8], [16]

It is apparent that the definition contains health status description and analyzes its causes. Outcomes of the study are applied to the solution of health and healthcare issues. We referred to the fact that a healthcare specialist must be aware of the occurrence of conditions that either improve or impair health. Not all these facts are included in the area of epidemiology interest; thus we state such facts that describe characteristics of healthcare services aimed at protecting/ restoring health, as well as such ones with indirect effects – mainly economic factors.

Public health information system (IS) in EU

Decision of the European Parliament and European Council No. 1786/2002/ES dated Sept 23, 2002 on the Action Plan of EC in the area of public health (2003 – 2008) states the obligation to prepare statistical section of the Public Health IS in cooperation with EU member countries, using the EC program as required, in order to support synergy and avoid duplicity. Decision of the European Parliament and European Council No. 1350/2007/ES dated Oct 23, 2007, constituting the 2nd Action Plan of EC in the area of public health (2008 – 2013) states that it is aimed at creating and spreading information and knowledge in the area of health, which should be reached through the activities oriented to further development of permanently sustainable health monitoring system, including the mechanisms for collection of comparable data and information with respective indicators, and to development of statistical parts of the system within the EC Statistical Program.

The regulation [17] constitutes the common frame of systematic creation of EC statistics in the area of public health and occupational safety and hygiene. The statistics shall be prepared in compliance with the principles of impartiality, reliability, neutrality, cost effectiveness and confidential nature of statistical data. In the form of harmonized and common system of data, they include the information required for EC activities in the area of public health, to support national strategies of the development of high quality, generally available and permanently sustainable healthcare, and for EC activities in the area of occupational safety and hygiene.

The document also sets forth quality requirements for data supplied to EC by EU member countries. Data quality attributes are as follows:

- „relevancy“, related to the extent in which the statistics meets current and potential needs of users;
- „accuracy“, related to the extent in which the estimation approximates unknown actual values;
- „timeliness“, expressing the time gap between certain event and availability of information describing the event;
- „time accuracy“, presenting the time delay between the data publishing date and target date of data supply;
- „availability“ and „transparency“, related to the terms of procurement applicable to data obtaining, use and interpretation by the users;
- „comparability“, related to the extent of reach/ effect of the difference between used statistical terms and measuring tools and practices applied to the comparison of statistics within geographical zones, sector areas or within specified time;
- „coherence“, related to data reasonability, the data can be combined by two manners and for various purposes.

The above stated features shall be considered not only when working with data obtained from domestic sources but also those of foreign origin. It is necessary to point out that Slovakia, as a member of European Community (EC), is committed to observe these principles and subject to the data quality audit conducted every 5 years. This contributes to increased data quality in Slovakia and the whole EU Community.

Information required for development of strategies

Organization and management of the system of health development, support and assurance requires detail knowledge of the system and its surroundings behavior. It is not necessary to ascertain ourselves that availability and understanding of quantitative, i.e. numerical characteristics of the system represents an inevitable precondition of good management along with information on legal, cultural, value and political environment. Such knowledge is necessary for answering a broad spectrum of questions, including these: Is the population in my surroundings (enter community, town, city, region, district, country…) healthy? What health problems should be addressed immediately? Do we have sufficient tools to solve them? The questions continue after answering them. We further ask: Was the applied procedure actually efficient? Did all who required the service (enter diagnostics, therapy, vaccination, transfer, etc.) actually receive it? Questions simplifying answers regarding selection of the best practice also belong here, for example: Which vaccine is the most appropriate in particular situation? Which of the available approaches has the fewest side effects?

Health strategy and policy

Seeking answers to these questions represents the process of decisions making on basis of situation identification, its analysis and interpretation. Traditionally, such decisions used to be made by physicians on various management levels. Such approach is unsustainable today in ever more complicated situation and increasing cost of the public health services. Accordingly, more and more problems have become the public policy topics in the form of regulations, training programs, funding or direct service provision. This publication has set forth the goal to show the way to students and experts towards obtaining knowledge and skills required for high quality decision making and assessment of adopted measures efficiency. We call these processes a strategy, since the Slovak language doesn´t distinguish between the politics of political parties and policy of the public health, or other target behaviors. In the management, strategy is defined as a "determination of basic long-term corporate objectives, methods of their fulfillment and allocation of funds required for fulfillment of the objectives". [18] Definition of policy is more complicated if it isn´t exclusively a political party matter and is called "policy" in English-Saxon environment (contrary to "politics"). It´s often said that it is a contextual, factual measure of policy dealing with the content and consequences of the state- political decisions requiring professional knowledge in particular specializations. While strategy shows where we want to go to, policy identifies the best way how to get there. In this publication, we will refer to strategy in case of both processes.

Why we use {R} environment

Free availability of {R} environment represents one of major advantages for students or researchers with limited finances. Another advantage refers to fast innovation of packages serving to various purposes. Relative disadvantage refers to austerity and limitations resulting from simple user environment. It is not a "clicking" program but requires certain time for studying it in order to master its effective application. Demanding first steps are compensated with possibility to repeat similar or equal procedures, which saves time and efforts. It is also a continuously increasing spectrum of offered solutions. Sufficient amount of manuals and books for the beginners and intermediates is available nowadays, describing application of the environment to statistical purposes. We will discuss it in detail in a separate chapter. Reader should not be discouraged but try to work in {R} through the examples stated herein.

Conventions use in tables and text

Each chapter starts with a brief introduction of the following text. Summary is stated at the end of each chapter, as well as the references.

Scripts in {R} are written in italics. These mean orders expressed through the agreed words or signs representing the manual on how to operate the programs in {R}. Script in the tables and on the pictures is also written in italics. However, results of script activities are written in regular font. Italics is also used to highlight the cited wording like in the original document. It is mainly done in order to emphasize the document originality. Some paragraphs are marked with headline at the beginning in order to make it easier for a reader.

Documents are often cited herein but mainly English documents are translated freely in order to present as close meaning as possible to the reader. We use terminology that corresponds to commonly used Slovak terminology as much as possible, if any. If it was impossible to find a synonym in Slovak language, we chose either the Czech translation or a term that, to our opinion, matched best with the English meaning.

Regarding the names of pictures, we often use word "picture" when we intend to present a structure, concept or effect. Used term "table" refers to the lists, definitions and other verbal descriptions.

Reference is cited in the form prescribed by the British Medical Journal (Vancouver format) since the program supporting work with references doesn´t contain Slovak citation standard.

Summary

This chapter´s aim was to acquaint a reader with the book content and specialization. We explained the basic concepts and the tool applied for demonstration of data processing. In the introduction, we explained the difference between the data and information, illustrating the meaning of this publication. We also presented the principles that we followed during the text creation. We would like to introduce the authors as well.

About the authors

The authors´ team consists of prof. MUDr. Martin Rusnák, CSc. and his wife prof.MUDr. Viera Rusnáková, CSc., MBA; both of them dedicated his professional career to the medicine and healthcare research, and perform as lecturers of statistical methods for physicians, students of medicine, public healthcare system and related specializations. Another team member, Mgr. Marek Psota was an intern at the Department of Public Healthcare, the University Trnava, at the time of the publication writing. Since he deals with disease modeling and processing of data from the available statistics, he covered the subject matters also in this book.

Prof. MUDr. Martin Rusnák, CSc. is a physician with professional specialization in the public healthcare, its quality, research of healthcare systems, medical and healthcare informatics and statistics, modeling, project management, healthcare and hospital IS. His scientific interest has been focused on the research of effectiveness of healthcare systems on population groups´ health in the area of serious brain injuries, in the area of minorities and population groups facing complicated economic and social life conditions. He often performed abroad and projects implemented together with the International Institute of Applied System Analysis are of major importance for this publication. In the institute, he had opportunity to work together with prestigious scientists in the area of demography and healthcare systems, for example prof. Nathan Keyfitz, recognized demographist often cited in respective chapters herein. He became familiar with prof. Dennis Meadows, known for his prediction called *Limits to Growth*. He witnessed startup of the program WHO Cindi, where he worked on the disease rate estimation through the models, based on the death rate, as a member of the research team at the University Heidelberg. He performs as a pedagogue and researcher at the Department of Public Healthcare, Faculty of Healthcare and Social Work, the University Trnava. Before he performed as a president of the International Neurotrauma Research Organization in Vienna, he had focused on the brain injury research. He is an author of more than 100 scientific and professional publications.

After passed specialization study about the medical informatics and clinical specialization neurology, prof. MUDr Viera Rusnáková, CSc., MBA dedicated her professional career to healthcare informatics and application of expert systems in the neurology. Later she focused on healthcare and quality management, along with teaching students and healthcare staff in the area of informatics, bio-statistics and public healthcare. In 2000, she was awarded with the title MBA upon graduation at the University Leeds, United Kingdom. She coordinated many international projects of healthcare management development in Slovakia, implemented together with US partners within EU. The core of her current research and pedagogic activities is in the area of statistics, safety and quality of healthcare. She is a coordinator of the project of performance and quality indicators in PATH hospitals in Slovakia.

Mgr. Marek Psota is a graduate of the Faculty of Healthcare and Social Work, University Trnava in Trnava, specialization Public Healthcare.

During his performance at the Department of Public Healthcare, he has focused on application of epidemiological methods primarily in the area of chronic diseases. During the study, he participated on a few national and international projects. At the Faculty of Healthcare and Social Work, he participated on teaching of statistics and data analysis through{R}, epidemiology of chronic diseases and intervention studies. In this publication, he was a co-author of the chapter dealing with summary health indicators and prepared the chapter about application of mathematical models on the public health survey.

References:

1. Makovický, E., Mariányi, J., Ďurian, J.: Social Medicine and Healthcare System Organization. Vol. 2. reworked and completed issue from 1981: Osveta, Martin.
2. Ashby, W. R. *An Introduction to Cybernetics*. 1956, London: Chapman & Hall.
3. *Wiener, N. Cybernetics: or Control and Communication in the Animal and the Machine. Second ed. 1965, Massachussets: Massachussets Institute of Technology*.
4. Spasoff, R. A. *Epidemiologic Methods for Health Policy*. 1999: Oxford University Press Inc. 240.
5. Kalibová, K. *ABC of Demographics*. 2 ed. 2001, Univerzita Karlova, Praha: Karolinum.
6. Field, M. J., Gold, M. R. ed. *Summarizing Population Health: Directions for the Development and Application of Population Metrics*. 1998, national Academies Press: Washington, D.C.
7. Tulchinski, T. H., Varavikova, E. A. *The New Public Health. 2 nd Edition*. 2nd ed. 2008: Academic Press. 696.
8. Porta, M. *A Dictionary of Epidemiology*, in *A Dictionary of Epidemiology*, M. Porta, S. Greenland, and J.M. Last, Editors. 2008, Oxford University Press, New York: Oxford. p. 289.
9. Parrish, R. G., Friedman, D. J., Hunter, E. L. *Defining Health Statistics and Their Scope: Shaping Policy and Practice to Improve the Population's Health*. In: *Health Statistics*, Friedman, D. J. Hunter, E. L., Parish, R. G., Editor. 2005, Oxford University Press: Oxford. p. 536.
10. Last, J. M. *A Dictionary of Epidemiology.* 1995, Oxford: Oxford University Press. 196.
11. Last, J. M. *Vocabulary of Epidemiology,* 1999, Bratislava: USAID.
12. Buehler, J. W., Hopkins, R. S., Overhage, J. M., Sosin, D. M. T. *Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group.* MMWR Recomm Rep, 2004. **53**(Rr-5): p. 1 – 11.
13. O'Carroll, P. W. *Introduction to Public Health Informatics*. In: *Public Health Informatics and Information Systems*, O'Carroll, P. W., Yasnoff, W. A., Ward, M. E., Ripp, L. H., Martin, E. L. Editor. 2010, Springer Verlag: New York. p. 751.
14. Brownson, R. C., Fielding, J. E. and Maylahn, C. M. *Evidence-based public health: a fundamental concept for public health practice.* Annu Rev Public Health, 2009. **30**: p. 175-201.
15. SR, Š.ú. *Demographics.* 2013 [cited 2013 28. 08.]; Available from: http://portal.statistics.sk/showdoc.do?docid=2434.
16. Last, J. M. *Making the Dictionary of Epidemiology.* Int J Epidemiol, 1996. **25**(5): p. 1098-101.
17. EU. *EU PARLIAMENT AND COUNCIL REGULATION on EC Statistics in the Area of Public Health and Occupational Safety and Hygiene.* In *1338/2008*, EP/ EC, Editor. 2008, EU: Brussels.
18. Chandler, A. D. *Strategy and Structure: Chapters in the History of the American Industrial Enterprise*. 2003: Beard Books.

Chapter 2

Data Preparation and Processing, Introduction to {R}

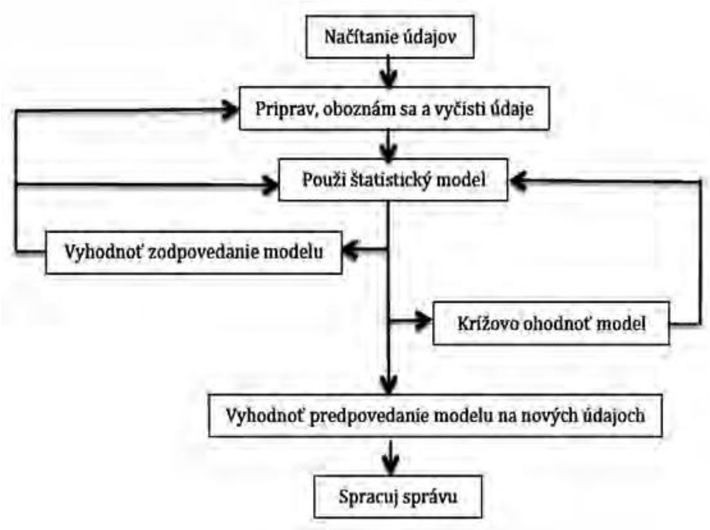
Chapter objectives

In this chapter, we will describe three basic steps of data preparation and application of the {R} environment tools, Excel, and DB functions. Data processing methods mentioned in the introduction and frequently used when working with data, characterize population health condition or services related thereto. Data collecting steps as well as frequently used questionnaire method are described in detail in the book about bio-statistics. [1] Here we want to deal with questions related to data processing from official sources, mainly from national or international statistics.

This chapter is aimed at providing a reader with the overview of the project {R} environment tools that will enable him to work with both simple and more complicated data structures in a purposely way, using as broad possibilities as possible, offered by {R}. Their application requires certain extent of imagination, since the data structure is best visualized on the surface or in the space. Even if the authors prefer to avoid formal mathematical expressions, two- and multi-measure data structures work with indexes and mathematical expression is usually most suitable to work instruction recording.

Data processing steps

Sequence of steps during data processing differs depending on the source. In this publication, we will focus on those collected and processed by an institution dealing with documentation of development in any of the health areas. Therefore we didn´t pay attention to primary data preparation, since our primary data refer to the numbers coming from routine collection of statistical data. We presume that data have been cleaned and are pre-processed, verified and validated. The first step then refers to data downloading from the existing files (currently most frequently available on the Internet) (Picture No. 1).



Picture No. 1 Data processing steps pursuant to [2]

Downloading

When working in the project environment {R}, data downloading from the file is facilitated by functions. We will describe them in detail in this chapter as soon as we become familiar with {R} fundamentals and data structures. We would only repeat here that the data files can be in various formats; those most frequently used include DB files, Excel, text files with separators, e.g. comma and tabulator.

Transformation

Transformation is usually required during work with data of certain measure. For example: conversion between units from other than metric system or conversion of currencies, or combination of two or more variables to the new one, for example calculation of the Body Mass Index. Below mentioned data processing methods will be used here. Work with the date is a specific data transformation example. We often used to get the age from the birth date or calculate duration of an event in years, months, weeks and days. These procedures will be described in the following sub-chapters.

Overview

Analysis should start with concise data study. Presuming that the entry data come from verified source, we don´t have to be afraid of the quality but it is appropriate to take in account the source-related limitations when interpreting the data. For example, when processing the data related to reported diseases (both infectious and non-infectious) we should consider that not all cases were reported for whatever reasons. For example, gonorrhea is not always reported since the patient doesn´t wish it for societal reasons. Many cases of influenza remain unreported since the patient treats himself and doesn´t draw a sick leave. It has been proved that sick leave drawing strongly depends on the income level and people with higher income tend to treat themselves without skipping worktime.

Data overview is best provided in the form of tables, illustration through scatterplot, histogram or other visualization form. In compliance with Rothman et al. [3], we would like to point out frequently misleading statistical parameters that automatically offer some statistical programs (p-value, certainty intervals). In the stage of orientation with and decision-making on further process, it could improperly affect the student´s decision. Concurrently, detail study of the original data could indicate trend analysis or seasonality methods, or imply hypotheses that can be later confirmed or rejected. Last but not least, such study could reveal some errors skipped by the original data processing subjects; however this is a very rare case when working with data from professional sources.

Statistical analysis

Statistical testing of hypotheses and estimations should be made with judgment and consideration of eventual systematic bias. Conclusions should be drawn after removal of as many potential distortion sources as possible, applying data standardization or other uni- or multi-variant procedures. Good knowledge of chosen statistical inference method (deduction) or model represents a precondition of correct conclusion formulation. Care should be taken when making judgment about correlations or effects of the variables with one another. Taking in account possible error (of I. or II. type) makes a researcher to draw likely interpretation, avoiding him to make imperative statements about the effect presence or absence.

Report

While the publication is not aimed at dealing with reports writing about the facts observed, we consider important to make some remarks anyway. At first, we would like to emphasize the need for reports formulations about the research outcomes in the simplest and best understandable manner. Clear definition of research objectives based on the global knowledge status represents the basic precondition of a successful report (we consider successful the report that brings a new finding). Such definition of the objectives usually results in proper application of the research methods, high quality presentation of the outcomes and finally also factual discussion. In Slovakia, we definitely don´t recommend extended structure of reports to theoretical and practical section since it impairs logical sequence of structured report and induces unreasonable and meaningless piling of references (in better case; or counterfeiting in worse case).

Environment {R}

Terminology

Since all examples were made through recalling the functions written for environment {R}, it is necessary to enlist general procedures typical for the environment. At first, we should make clear the basic words, i.e. terminology.

Many users consider {R}a statistical program. It is not false but we would like to state that {R} is a program environment (for development of computer programs), in which functions are developed or used for application mostly in the statistics. [4] Thus it contains also data packages, enables plotting of data or results of solutions, creating the tables and cooperating with other program wholes, e.g. with Excel. We can state that the environment is able to sufficiently substitute historically older, many years developed professional statistical programs, for example SPSS, SAS and others.

Functions

Function represents a program recalled from the master program, in this case from {R} environment. Functions process parameters and return the processing results in the form of values. General format of function recalling consists of the function name and parameters in the brackets: *function (parameter, ...)*. To make them well-arranged and purposely, the functions are grouped in libraries. Library contains many functions that are installed at once in the program environment and functions can be recalled one after another. Libraries are freely available in the form of packages in CRAn. This abbreviation refers to the network of servers containing the same set of sources (programs and packages) of {R} environment (http://cran.r-project.org). The manual [5]
contains the overview of basic functions and definitions of their structure.

Variable

Variable is a term often used in the statistics and work with numbers. Using calculator, we enter a number and operator, then another number and press "equals", getting the result. However, we cannot recall the result after switching the calculator off and then on, entering another calculation. In this case, variable is a solution. If we name the result in PC, instead of in the calculator, for example in the program {R}, the program will remember the number under that name till the end of the program, or until its storage on the disc. This is an advantage of using variable. Its name refers to the fact that various data can be hidden behind the name of a single variable. Variable can contain one or more numbers that we can obtain mostly through measurement of a unit, thus it is called *measurement.* Typical object of statistical survey refers to a set of variables that contain various measurements. For example, searching for impact of nutrition on the growth of a child, we will get the variables that include measurement of the children – age, gender, total intake in kcal per a day, amount of fat, proteins and carbohydrates in the consumed food, etc. All data entered for example in Excel shall be stored on PC disc in the file called *child,* and thus save the variables for later use.

Installation of program environment {R}

Project website http://www.R-project.org/ is the source of program{R}. Despite of program availability for various operating systems, we shall focus on Windows. If any user has a different operating system, e.g. Linux or MAC OS, he should download the required version on the mentioned website. Process of downloading and installation is described in the table No. 1.

- Go to project website http://www.R-project.org/
- Click on the link CRAN in section Download on the left site column
- Choose the website you will download the program from
- Click on Windows in section Download and Install R
- On the following page, click on "base"
- Activate the link Download R 2.11.1 for Windows (33 megabytes, 32 bits)
- After downloading, unpack the program and install it. Successful installation ends with displayed blue letter R on the monitor display

Table No. 1 {R}program installation procedure

Objects

All we work with in the environment {R} is called *object*, i.e. signs, numbers, functions, etc. Working with objects, we use the names allocated to the objects, expressing the object value. We distinguish a few types of objects; they are classified and called *classes.*

- numbers
  - integers – 2, 3, 4
  - double – .5, 3.8, 4.3
  - complex – 4+2i, 25-3i, i;
- characters – “test”, “Hi”;
- date – more complicated assignment but the object looks as follows “2006-03-15”;
- logical values – TRUE, FALSE;
- vectors – number, chain or other data strings,
- factor – special vector type with nominal or ordinal data,
- matrix – classical mathematical matrix or matrix of data and chains,
- array – multi-measure matrix;
- lists – specific objects grouping various other objects of different types;
- data frames – classical DB tables;
- functions – orders that activate certain operation with objects;
- expressions, formula, special objects (test, analyses results).

Picture No. 2 Object classes pursuant to [6]

When allocating names to objects, it is appropriate and sometimes necessary to follow certain principles. Names of functions and certain values that are reserved shall not be used, e.g. a variable cannot be marked as *F*, since it is the sign with meaning *FALSE*. The system will warn you on error if you violate this principle. Similarly, it is necessary to strictly distinguish small and big letters. Objects named *first* and *First* are not equal but refer to two different objects that could have different content allocated. Terms containing two or more words shall not contain gap or separator; the words can be joined through dot or break character. For example, *first name* as an object name is incorrect but *first.name* or *first_name* is acceptable.

```
>first name <- c(„František“) # system signalizes error
Error: unexpected symbol in “first name”

>first_name <- c(“František”)
>first_name
[1] “František”

>first.name  <- c(“František”)
>first.name
[1] “František”
```

Example No. 1 Correct and incorrect names of objects

Certain functions require strict observance of the object class. If we are not sure about the class, e.g. after reading external data, we can apply some of the functions *class* (), *mode*(), or *type of()*. Usually we can take the first one; further two provide less detail classification.

Certain functions facilitate work with objects. Here we will enlist only those most frequently used. For example, we need to identify available objects at the moment. Function *ls()* enlists all objects that are available at workspace at the moment. Function *objects()* has similar effect.

Function *rm()* serves for removal of object/s from the workspace. Names of objects we wish to remove shall be written in the brackets as a function argument. Function *str()* displays the object structure. We will describe the functions in detail during their practical use.

Simple operations with data and variables

In the previous text, we mentioned the term “variable” on the example of calculator limitations. {R} can be used as a calculator very easily; example No. 2 shows basic operations. Sign # means comments that are ignored by program {R} but later it allows for reminder of certain steps intention.

```
>3+4 # sum
[1] 7
>5.6-3.2 # difference
[1] 2.4
>4.5*3          #
multiplying     [1]
13.5
>9.6/4       #
dividing [1] 2.4
```

Example No. 2 Basic operations with numbers

If we wish to save result of certain operation, we have to save it as a variable. The easiest way how to do it is use of sign “equals” = (Example No. 3).

```
> a = 3*4 # we multiplied two numbers and stored the result in variable a
> a # result is written upon entered name of
variable
> [1] 12
> b=6^2 # operation of square
> b # result is written upon entered name of
variable [1] 36
> ls() # function writing names of variables stored in the memory of program {R}
[1] "a"                    "b"
> a*b # we can treat variables similar to numbers [1] 432
> b-a
[1] 24
> result<-a+b # result can be stored in a new variable
> result
[1] 48
```

Example No. 3 Use of variable for storage of operation results

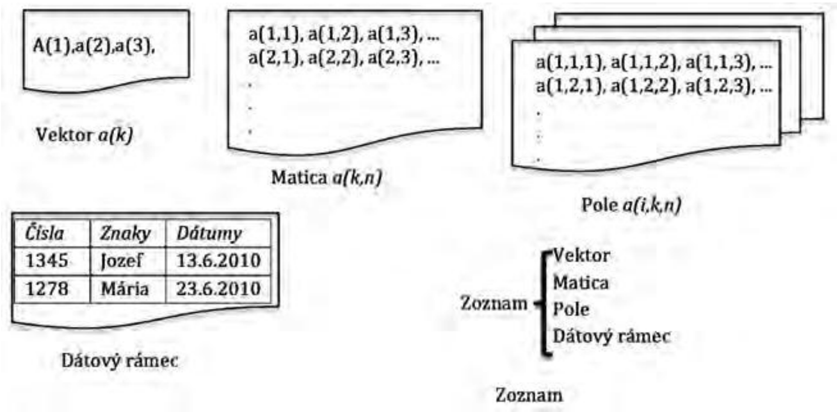
Variables can contain whole and decimal numbers, as well as letters and signs. Moreover, variables can contain logical values *YES* and *NO*. Missing value NA is a special case having important status in the statistics. Letters and texts are given in the quotation marks and can be stored in variables. Logical variables should be entered either with big letter or with the first letter only (Y and N). Missing value shall be marked with combination of letters NA (Example No. 4).

```
# allocation of texts to variable
> texts <- c("Martin", "Juraj", "Zdeno", "Karol")
> texts
[1] "Martin" "Juraj" "Zdeno" "Karol"
> texts <- c("Martin", "Juraj", NA, "Zdeno", "Karol") # missing value
> texts
[1] "Martin" "Juraj" NA  "Zdeno" "Karol"
# logical variables
> yes <- T
> yes
[1] TRUE
> yes = TRUE
> yes
[1] TRUE
> no<- FALSE
> no
[1] False
> no <- F
> no
[1] False
```

Example No. 4 Creation of variables with various contents

Data structures

Along with variables with the only number, we use more complicated variables, e.g. vectors, matrices, arrays, data frame and lists (Picture No. 3). They differ in data type variability, structure complicatedness and method of particular items availability.



Picture No. 3 Various types of data structures

We will gradually describe all of them.

Vectors

Vectors are often used in statistics. They contain measurements of various units from one variable. We entered the variable "age" containing the age of children. To enter values, we used function *c()* that combines or allocates values in its argument and stores it in a single variable in the form of vector (Example No. 5).

```
>age<- c(1, 2.5, 2, 4.5, 3, 6)
>age
[1] 1.0 2.5 2.0 4.5 3.0 6.0
```

Example No. 5 Creation of measurement vector and its storage in variable

Results is similar in case of used < – or = for allocation. Names of variables are not limited in any way but one should take care of small and big letters. As well, avoid letters with diacritics; {R} environment can accept some of them but not all.

Matrices

Creation of matrices often used for statistical problems solving is more complicated. Matrix is a table containing a few variables, each of them with many measurements. Matrix size is defined by number of columns and rows, together called *measure*. Matrix creation from vector of values requires addition of matrix measure, i.e. specification of number of rows and columns through function *dim()*. We will demonstrate its use on the above created vector (Example 6).

```
> age
[1] 1.0 2.5 2.0 4.5 3.0 6.0
> dim(age) <- c(2,3) # we created matrix with 2 rows and 3 columns from vector with
6 measurements
> age
   [,1] [,2] [,3]
[1,]  1.0  2.0  3
[2,] 2.5 4.5 6
> dim(age) <- c(3,2) # we created matrix with 3 rows and 2 columns from vector with
6 measurements
> age
   [,1]   [,2]
[1,] 1.0 4.5
[2,] 2.5 3.0
[3,] 2.0 6.0
```

Example 6 Creation of matrix from vector of numbers

Use of function *matrix()* is more elegant and broader, providing possibilities of the data matrix creation upon entered vector of values that are then arranged in columns and rows, whose names can be specified later. Function entering format:

matrix < – matrix(vector, nrow=number_of_rows, ncol=number_of_columns, byrow=logical_value, dimnames=list(char_vector_rownames, char_vector_colnames))

Selection *byrow=TRUE* determines matrix filling in rows, *byrow=FALSE* determines matrix filling in columns. Preselected value is in columns. We will demonstrate some operations on the example of the above used vector.

```
> age < – c(1, 2.5, 2, 4.5, 3, 6) # creation of vector of numbers
> age1 < – matrix(vek, nrow=3, ncol=2) # vector is transformed to matrix with 3 rows
and 2 columns; notice that the function would use pre-selected value byrow=FALSE
without determination of parameter byraw and process the request in columns.
> age1   #    result
  [,1] [,2]
[1,] 1.0 4.5
```

[2,] 2.5 3.0

[3,] 2.0 6.0

> *age2 < – matrix(age, nrow=2, ncol=3)* # at 2 rows and 3 columns but maintained sequence in columns

> *age2* # result

[,1] [,2] [,3]

[1,] 1.0 2.0 3

[2,] 2.5 4.5 6

> *age1 < – matrix(age, nrow=3, ncol=2, dimnames = list(c(„A“,“B“,“C“), c(„X“,“Z“)))* # we added the names of columns and rows

> *age1* # result

X Z

A 1.0 4.5

B 2.5 3.0

C 2.0 6.0

> *age2 < – matrix(age, nrow=2, ncol=3, byrow=TRUE)* # changed loading from columns to rows

> *age2* # result

[,1] [,2] [,3]

[1,] 1.0 2.5 2

[2,] 4.5 3.0 6

> *age3 < – matrix(age, nrow=2, ncol=3, byrow=FALSE)* # preset

> *age3* # result

[,1] [,2] [,3]

[1,] 1.0 2.0 3

[2,] 2.5 4.5 6

Example 7 application of function *matrix* ()

Matrices represent a useful method of data processing but we have to know correct identification of particular values. We proceed similar to any other tables; column and row intersection clearly defines the number position. We mark it with indexes, for example *A [i,]* marks *i*-row of matrix A, while *A [,j]* corresponds to *j*-column of the same matrix. Joining it together in a single entry, then *A [i, j]* gradually corresponds to *ij*-number. These indexes *i* and *j* can be also in the form of vectors so as they enable selection of a few rows and columns at once.

> *age<- c(1, 2.5, 2, 4.5, 3, 6)*

> *age2 <- matrix(age, nrow=2, ncol=3)* # we create data matrix from the above data with 2 rows and 3 columns

> *age2* # result
[,1] [,2] [,3]
[1,] 1.0 2.0 3
[2,] 2.5 4.5 6

> *age2[2,]* # we choose only 2nd row
[1] 2.5 4.5 6.0

> *age2[,2]* # we choose only 2nd column [1] 2.0 4.5

> *age2[,c(2,3)]* # we choose only 2nd and 3rd column
[,1] [,2]
[1,] 2.0 3
[2,] 4.5 6

Example 8 Selection of numbers from matrix

Matrices have two measures and can contain only single-type data, similar to vectors. If we have more than two measures, we should use arrays. If we need to combine various data types, we should use data frames.

Arrays

Arrays are matrices with more than two measures. We can imagine them as matrices arranged after each other. We state them despite of not commonly used methods described in this publication. Therefore we use the example of Kabacoff [2], using array creation function *(vector, measures, dimnames)*. Variable *vector* contains data in the form of vector, *measures* represent another vector determining the biggest index from each of the measures and *dimnames* is optional list of measures description. The following Picture No. illustrates creation of 3D array of (2x3x4) numbers.

> *dim1 <- c(„A1“, „A2“)*
> *dim2 <- c(“B1”, “B2”, “B3”)*
> *dim3 <- c(“C1”, “C2”, “C3”, “C4”)*
> *z <- array(1:24, c(2, 3, 4), dimnames=list(dim1, dim2, dim3))* # vector is created from numbers within 1 – 24, the numbers are divided in 2 rows each of 3 columns and 4 measures; rows are marked with letter A, columns with letter B and each of the measures with letter C
> *z* # enter each of the measures
, , C1
B1 B2 B3
A1 1 3 5
A2 2 4 6

```
, , C2
  B1 B2 B3
A1 7 9 11
A2 8 10 12

, , C3
  B1 B2 B3
A1 13 15 17
A2 14 16 18

, , C4
  B1 B2 B3
A1 19 21 23
A2 20 22 24
```

Example 9 Example of data array creation with function array() pursuant to [2]

Data frame

Data frames represent most frequently used structure for work with data. You can imagine data frame as a matrix where each column can contain other data format. Analogical to the previous frame creation method (skip word „data" to make it easier) we shall use function *data.frame(col1, col2, col3,…)*.

As an example, imagine creation of file with particular hospital wards, each with allocated number of beds, physicians and the accreditation status. Process of such file creation is illustrated in example No. 10.

```
> ward <- c(„internal", „neurology", „surgery", „pediatrics") # variable with text
> beds <- c(80, 25, 60, 50) # variable with numbers
> accreditation <- c(TRUE, FALSE, FALSE, TRUE) # variable with logical values
> hospital <- data.frame(ward, beds, accreditation)
> hospital

  ward           beds  accreditation
1 internal       80      TRUE
2 neurology      25      False
3 surgery        60      False
4 pediatrics     50      TRUE
```

Example 10 Creation of data frame

Creation of data frame is by far not complicated. However, how to address particular data cells? Entry via indexes is the easiest way, similar to matrices.

This method is simple but tedious, with high potential of errors (Example No. 11).

```
>hospital[1,3] # data cell determination according to row
and column No.[1] TRUE

>hospital[,2] # all values from column 2
[1] 80 25 60 50

>hospital[2,] # all values from row 2
   ward beds accreditation
2 neurology  25   False

> hospital[2:3] # all values of column 2 and 3 (preset column selection)
         beds   accreditation
 1       80       TRUE
 2       25       False
 3       60       False
 4       50       TRUE
```

Example 11 Selection of element from data frame through direct address entering

Other method refers to column specification based on its name. Sign *$* separates the frame name from the column name; thus in case of column 2 named *beds* it shall be specified with *hospital$beds.* If we wish to choose only 3rd item in the column, we shall add a row *hospital$postele [3]* with resulting No. 60.

This method is neither a good idea since we need to write down entire combination of frame and column name in case of every operation and it is rather impractical. {R} environment offers many functions that simplify these operations significantly: *attach()*, *detach()* and *with()*. The first one of them, *attach()*, adds the names of variables in the {R} environment search that makes the columns looking like independent variable and can be treated through the names.

```
> attach(hospital)
The following object(s) are masked _by_ '.GlobalEnv':

  accreditation, ward, beds

> beds
[1] 80 25 60 50
```

Example 12 Use of function *attach()*

Cancellation of the function effect, namely removal of the names of variables from the {R} environment searching can be done through function *detach()*. This makes names available for variables in other arrangements.

Despite of not using as many variables as necessary, use of this function is recommended at completion of work with the variables from one data array.

Instead of writing *structure$variable,* we can use function *with()*. For example, entering structure and variable in the function argument will display the variable values (Example 13).

```
> with(hospital, beds)
[1] 80 25 60 50
```

Example 13 Use of function *with()*

Even if this function use seems not necessary, it often facilitates work in cases of one-time variable application in more complicated functions.

The list refers to a specific data structure enabling combination of various data structures under single name, for example vectors with matrices and data frames. It is most frequently used when returning values from recalled function.

Date and time

Date and time is often stated in the records, be it the date of any event, birth or death date. Working with variables containing time is not trivial, thus we discuss it here; however, the time as a variable is not considered during work with already pre-processed data.

Let´s start with current date that is uploaded by {R} from the date monitored by PC operating system. If you get the date different from the current one, don´t blame {R} immediately but look at your system time preset in PC.

```
> date()
[1] “Sun Jan 6 13:15:41 2013”
> Sys.time()
[1] “2013-01-06 17:24:46 CET”
> Sys.Date()
[1] “2013-01-06”
```

Example 14 Time of used system recalled by three functions

Date and time can be allocated to the text-type variable either directly or we will work with it in a limited extent. Date is allocated through the allocation function *c()*. To perform other activities with date and time, we have to learn about two classes *POSIXct* and *POSIXlt*. [5] The first one - *POSIXct* represents number of seconds (with sign), that have passed since 1970 in the form of number vector. The latter - *POSIXlt* is a list of vectors representing the parts of date and time specification (Table No.2).

sec	Seconds 0–61
min	Minutes 0–59
hour	Hours 0–23
mday	Day in month 1–31
mon	Months after the 1st in year 0–11
year	Number of years since 1900
wday	Week day starting with Sunday 0–6
yday	Day in year 0–365
isdst	Summer/ winter time label. If used +, if not used 0, if unknown -

Table No. 2 Date and time specification in class *POSIXlt.*

To be able to use the list of vectors, we need a function that is able to insert it in a variable. It is function *as.POSIXlt()*. Calling it and applying the system time to its argument, we can break down the list to particular items, using the names of the list vectors in compliance with the above stated table (Table No. 2).

Examples of *as.POSIXlt(Sys.time()* function application to work with time are stated in the following example (Example No. 15).

```
> date<- as.POSIXlt(Sys.time()) # list with actual system time was entered in the
variable
> date$sec # second
[1] 22.47269
> date$min # minutes passed from current hour
[1] 11
> date$hour # hours passed from current day
[1] 18
> date$mday # days passed from current
year [1] 7
> date$mon # order No. of current month [1] 0
> date$year # order No. of current year starting from 1900 [1] 113
> date$wday # order No. of week day
[1] 1
> date$yday # order No of day in year
[1] 6
> date$isdst # summer/ winter time
[1] 0
```

Example 15 Breakdown of POSIXlt list items

Identification of time passed between two time points refers to frequently performed operation that can be simplified through function *difftime(time1, time2, tz = „ “, units = c(„auto“, „secs“, „mins“, „hours“, „days“, „weeks“))*. The first two variables refer to both time moments, be it date or hour. The label units determine selection of required results format (abbreviations can be used as well). Using this function, it is necessary to convert the result to number/ digit, especially if we want to continue with an arithmetical operation, for example difference in years, i.e. divide by number of days in a year (Example No. 16).

```
> today<- as.POSIXct(Sys.time()) #entered current system time
> here<- as.POSIXct("1992-03-25") # opening date of reconstructed University in Trnava
> diff <- difftime(today, here, units = c("d")) # number of days passed from the
  Opening Day
> diff
Time difference of 7596.618 days
> class(diff) # difference still in class diff
[1] "difftime"
> diff <- as.numeric(diff) # presented difference in the form of digit
> year <- diff/365 # number of years
> year
[1] 20.81265
```

Example No. 16 Difference of two dates

Time processing issue in the statistical systems is not so frequent to entitle us to deal with it in detail. If a reader faces the time problem solving, we recommend searching the solution either on the Internet or in the books about {R} [2, 7].

Data reading from files

The easiest way of file reading is to present it for example in the program Excel or Word, or via Adobe. Then we take the required data in a block and copy them through CTRL-C or the toolbar order. To read the data from the clipboard, use the function read.table*()*. In the first step, prepare the data file you wish to burden. For demonstration, we will use the file in Excel (you can test it with your data), obtained as a freely available data file from the study about heart disorder risk factors in Los Angeles.[3] The data are in format Excel; we will store them in the folder in the format, where they are separated with comma*(.csv*).

[3] http://www.umass.edu/statdata/statdata/data/laheart.txt

It is caused by the existence of more formats of Excel files and the data entry requires the easiest possible way of reading. To simplify the data usage in this folder, we will specify it as a working call of the function *setwd()*. Its argument shall be the path that will determine folder storage on the disc. To verify the correctness, we use the function *getwd()* and the list of files contained therein shall be displayed with function *list.files()*.

Recalling function *read.table()* enables reading of the file and insert the results in the variable. The result refers to a variable in the form of matrix. If we want to work with data in particular columns, we should recall them with sign $ in the form *matrix$column*, thus in our case it will be for example addressing of the column with ID of survey participants and we shall use the structure *heart$ID*. Recalling function *attach()* and entering the variable name, we can use the names of columns as independent variables. Entire process is illustrated in Example No. 17.

```
> setwd(„C:/Users/Martin Rusnak/Documents/R/Datafiles/Cardio“) # work folder specification

> getwd() # verification of correctness
[1] "C:/Users/Martin Rusnak/Documents/R/Datafiles/Cardio"

> list.files() # folder content record
[1] "~$A. HEART DATA (LAHEART.DAT).docx" "L.A. HEART DATA
(LAHEART. DAT).docx" "laheart.xls" "laheart.csv"

> heart <- read.table("laheart.csv",header=T, sep = ",") # reading of data separated by comma, columns´ names are in the first row

> heart
  ID AGE_50 MD_50 SBP_50 DBP_50 HT_50 WT_50 CHOL_50 SES CL_STATUS
MD_62 SBP_62 DBP_62 CHOL_62 WT_62 IHD_DX DEATH_YR DEATH
1  1   42  1  110  65  64  147  291  2    8  4  120  78  271  146  2   68  1
2  2   53  1  130  72  69  167  278  1    6  2  122  68  250  165  9   67  1
......
>heart$ID # enlists ID column contents
>attach(heart) # allows for making the column name also the name of variable
```

Example No. 17 reading of file exported from Excel

If a reader considers it complicated or he wants to work only with a part of data frame (in Excel or other program), recall of similar function can be used but enter specification *„clipboard“* instead of the file name. Proceed so that you remove a part of data and take it in a block together with the 1st row containing names of columns. Then press *CTRL-C* and copy it to the clipboard. Then, recalling function *read.table(„clipboard“, header=T)* with arguments *„clipboard“* , stating that the data are in clipboard, and *header=T*, stating that the names of columns are in the first row, we can load the data in the variable in {R} (Example No. 18).

```
> heart1 <- read.table(„clipboard“,header=T) # read file from clipboard
> heart1
     AGE_50   MD_50  SBP_50  DBP_50  HT_50   WT_50  CHOL_50
1      42       1     110      65     64      147     291
2      53       1     130      72     69      167     278
3      53       2     120      90     70      222     342
4      48       4     120      80     72      229     239
5      53       3     118      74     66      134     243
6      58       2     122      72     69      135     210
7      48       4     130      90     67      165     219
8      60       1     124      80     74      235     203
9      59       4     160     100     72      206     269
10     40       3     120      80     69      148     185
11     56       3     115      80     64      147     260
12     58       3     140      90     63      121     312
13     64       2     135      85     64      189     185
14     57       2     110      78     70      173     282
```

Example no. 18 Data reading from the file in clipboard

We will present more complicated approaches in the chapters dealing with particular data from files.

Auxiliary programs

There are a few auxiliary programs that facilitate work with program environment{R} through selection from the offer. Library Rcmdr represents one of the simplest auxiliary programs. The library must be installed from the toolbar program {R} *Packages-Install package(s)*. Later it can be recalled with function *library()*

> library(Rcmdr)

Environment R Commander is very simple and offers work simplification. Detail manual is included in the program.

Freely available program Tinn-R[4] offers more opportunities.

[4] http://www.sciviews.org/Tinn-R/

The program installation is easy and offers also text editor along with provided access to {R}. Excel addition program is useful for working with larger data files, which enables call directly in the function {R}. Data loading from Excel is not necessary in this case and functions can be recalled directly from Excel toolbar. This addition to Excel can be also downloaded for free.[5]

Summary

This chapter´s goal was to indicate the spectrum of possibilities offered by {R} environment for basic work with data. We introduced {R} environment and terminology used therein. We described the types of objects and some functions enabling or facilitating work with objects. As well, we mentioned a few programs facilitating work in {R} environment and interconnecting it with common environment Windows or other operating system.

At the end of this chapter, we would like to recommend some publications to all subjects interested that contain more details about the environment {R} than presented here. All of them are available via Internet. The first one is a manual in Czech language from Pavel Drozd: *Exercise from biostatistics. Introduction to work with software {R}* [6]. The book is intended for students and provides extensive information with many practical examples. The second book was written by Kateřina Konečná and is also in Czech language; providing instruction on basic processes in this environment. [8] Another book is in English language and is very instructive. The author Kabacoff [2] chose simple but very compendious presentation of fundamental work in the environment {R}. Last but not least, I recommend downloading the list of basic functions of the environment{R} [5] and consult it as often as possible; mainly during initial stages of acquainting with {R} environment.

References

1. Rusnák, M., Rusnáková, V., Majdan, M. *Bio-statistics for students of public healthcare system.* I. ed, 2010, Trnava: Typi Universitatis Tyrnaviensis, Publishing Agency of the University Trnava 216.
2. Kabacoff, R. I. *R in Action: Data analysis and graphics with R*. 2011: ManninG, Shelter Island.
3. Rothman, K. J., Greenland S and Lash TL. *Modern Epidemiology, 3rd Edition*. 3rd ed. 2008: Lippincott Williams & Wilkins. 738.
4. R Development Core Team. *R: A language and environment for statistical computing*. 2008, R Foundation for Statistical Computing: Vienna, Austria.

[5] http://cran.r-project.org/web/packages/RExcelInstaller/index.html

5. The R Development Core Team. *R: A Language and Environment for Statistical Computing. Reference Index*, In *R Foundation for Statistical Computing.* 2009.
6. Drozd, P. *Exercise from biostatistics. Introduction to work with software R*, P.f., Ostrava University, Editor. 2007: OSTRAVA p. 111.
7. Crawley, M. J. *The R Book*. 2007, Chichester: John Wiley & Sons Lt.
8. Konečná, K. *How to work with R language*. 2011: Masaryk University, Institute of Biostatistics and Analyses.

Chapter 3

Classification serving for health statistics

Chapter objectives
Classification and terminology of diseases and injuries
International classification of diseases
Systemized Medicine Nomenclature – clinical terms SNOMED-CT
Payment for diagnosis (so called DRG)
Coding quality
Summary
References

Chapter objectives

The public and professionals have always been interested in the overview of diseases and their causes within the population. The essence of these diseases had not been disclosed to them for a long time. Development of human cognition was associated with the development of health and health disorders cognition. Taking in account broad spectrum of disease and health disorders, as well as recording these effects in the groups of population and calling for opportunity to use the information on national and international level resulted in the need for uniform terminology and uniform classification of diseases and death causes. Reader will be acquainted with the history of these efforts and become familiar with the development of international classification of diseases and related systems. A separate part is dedicated to the classification of clinical units and procedures, including their application for purpose of payment for healthcare facilities outputs. At the end of the chapter, we will deal with the data quality and its impact on utilization of coded data.

Classification and terminology of diseases and injuries

Bertillon´s [1] classification of death causes became the first internationally recognized classification system at the time of its publishing, serving for recording and statistical purposes. Later classifications were gradually derived from it. Based on the agreement reached by all participating countries in 1900, the classification and its principles were revised in 10-year intervals in order to reflect the status of cognition and the needs of healthcare statistics. [2]

Classification of diseases

Classification represents a fundamental procedure of statistical survey of effects. Scientific research inevitably requires uniform definitions used for classification based on common and agreed characteristics. In case of the study about diseases and deaths, the classification has become an urgent condition for cognition of the dynamics of the effects within the population. Uniform classification can be based on various preconditions; somebody could prefer anatomical localization, the other one could prefer the process that resulted in disease or injury; we can also imagine classification based on pathological process etiology. What is then a basic criterion of classification? It is a purpose for which the classification should serve. This is the reason why we have so many classification systems of diseases and health disorders nowadays. We will talk about a few of them; starting with the one that has become worldwide used. Majority of the classification systems has been still in development I order to monitor the human cognition progress.

Terminology

The depth of medical cognition can be identified according to used terminology as well. In the past, all water retention conditions were called “hydrops”. Nowadays we distinguish this condition according to the cause; be it body circulation system insufficiency or a consequence of renal or liver insufficiency, or some of many other possible causes. To meet its purpose, terminology must be so compendious that it exactly characterizes status quo and it could be then recorded. Currently we distinguish between many pathological conditions, their causes, symptoms and the severity level. Terminology of diseases, health disorders and injuries is unified internationally only to some extent. Some nations refer to the Latin terminology; other ones use their own expressions often resulting from historical traditions. German-speaking countries have many more specific terms than countries with prevailing English-speaking people. In these countries, Latin term represents the basis but is often written and read according to local habits. The Slovak terminology of diseases often refers to the Latin terminology in combination with original Slovak words. In the terms of classification, any condition we want to classify should be clearly expressed and defined.

English language knows three terms for pathological condition. [3] They are: „disease“, „illness“ and „sickness“. Disease refers to a functional and structural disorder of human body resulting in clinical effects and symptoms, marked as deviation from standard. [4] „Illness“ represents subjective perception of a disease.

Disease can proceed without perception of any symptoms or limitations by the ill person. High blood pressure is an example. The third term „sickness“ is used as a response of the society to the disease in the terms of changed relation of an ill person to the others. We can presume that the Slovak language will manage to assign a specific meaning to the above terms in the future.

What is the difference between the terminology and classification in case of health and disease? We said that the terminology provides for exact denomination of every condition called “health disorder”. Many verbal descriptions are then created that are often ambiguous, too broad or a few names exist for a single condition; thus many synonyms and eponyms are used.[6] [5] For example, hypertension is called also high blood pressure. This ambiguity doesn´t represent a major problem in the description of clinical condition of an individual. However, such situation disallows for definite recording, classification and evaluation of effects within the studied population for purposes of statistical processing of diseases and injuries. Therefore, classification is aimed at reducing number of the terms so as they cover all required pathological conditions, if possible. If it is proved that a condition not classified per se at the moment, representing a part of another condition, occurs frequently or is important from other aspect, there will be a reason to classification expansion or revision. So far we have discussed only diseases and injuries. Classification in medical and healthcare statistics often relates to other effects as well, for example special procedures. They have been ever more important for the public health, mainly in the terms of the new public healthcare system. This system has been increasingly dealing with the issues of professional healthcare relation to the diseases prevention on all levels.

Brief history

Relation of the development of terminology and classification in the history of medicine has adequately enlightened their dependency on the cognition, and the goals. Mentioned historical events resulted from the compilation of the facts published in various books. [1, 2, 6, 7] History of the mortal statistics recognizes the „London Bills of Mortality“, published on regular basis since 1603 during the pest epidemics in London. They are associated with the names John Graunt and Sir William Petty, considered as founders of the epidemiological demographics. [8] In the first, irregularly released issues from 1570, they stated religion, from 1629 it was the death cause and then the age of the late person was sufficient as the only fact from the beginning of the 18th century.

---

6 Eponym is a term joining an object with real or fictitious person or location named. In the medicine, eponym most frequently refers to the name of the author of significant scientific fact or is created as paying tribute to a person who discovered or described it as first, or significantly contributed to its spreading and is tightly associated with it so that his/ her name has become a denomination of the object.

John Graunt acknowledged the need for publication when he wanted to estimate the age of the late persons. Therefore he summed up all deaths marked as “soor”, “spasms”, “rickets”, “teeth”, “worms”, prematurely born children”, “children with enlarged liver”, and added 50% of deaths thereto, marked as “variolla”, “small-pox”, “chicken-pox”, and “worms without spasms”. In this way, he reached 36 % death rate until the age of 6 years. Such classification was very rough. In the 18th century, Francois de Lacroix (known as Sauvages) published the book called *Nosologica Methodica*, which was followed by another one from recognized system botanist Carl von Linné, called Genera *Morborum*. Classification developed by William Cullen, published under name *Synopsis Nosologiae Methodicae*, had been used till the beginning of the 19th century. In the 18th century publications of tables containing the causes of death appeared also in other European countries. Jean Razoux from France published such mortality tables for the first time from the hospital in Nimes in 1767. In Denmark, they were published for Copenhagen since 1709 during major pest epidemic. In Berlin, it was since 1737. The British registration system was the first one that spread to the whole territory; followed by France, Paris and later other French cities and smaller towns from 1906. The Scandinavian countries had gradually developed the registration since the half of the 19th century. At the beginning of the 20th century, majority of European countries had functional system of death causes recording based on medical certificate.

William Farr (1807 – 1883), let´s say a clerk of the statistical office for English and Wales, considered the first medical statistical clerk, used the Cullen´s classification. Later he realized lack of medicine progress resulting from the classification usage, thus insufficient also for statistical purposes. In the Annual Report of the office, he wrote: “Terminology is as important for this study specialization as the weights and measures for physical sciences, and should be implemented immediately.” He continued to deal with classification and terminology and published his research outcomes on regular basis in the annual reports. The first international statistical congress in the Brussels in 1853 entrusted Dr. William Farr and Dr. Mark d‘Espine from Geneva with preparation of internationally applicable uniform classification of the death causes. At the following congress, the two men presented two various lists. Mr. Farr classified the diseases in five groups, distinguishing between epidemiological, constitutional, anatomically localized, genetic diseases and diseases occurring as a direct consequence of violence. In Paris, the terminology was developed by Philippe Pinel and it has been used since 1802, passing through nine revisions and reflecting the development in cognition, especially in the bacteriology. In 1891, the International Institute of Statistics empowered Mr. Jacques Bertillon, head of the Statistical Institute in Paris, with preparation of the draft new international classification. He presented the draft document in 1886. [7] The document consisted of 14 chapters (currently 21) and contained 203 items with exact definition. The classification was revised for the first time in 1890 and since then we know it under the name International Classification of Diseases (ICD). Currently we use the 10th revision.

International classification of diseases

Original version

Complete name of worldwide used classification refers to *"International Statistical Classification of Diseases and Related Health Problems 10th Revision, Version for 2007 (ICD)"*. [9] It was adopted at the 42nd General Assembly of the World Health Organization (WHO) in 1990 and has been used in WHO member countries since 1994. Traditional MDC classification was saved but the previous nomenclature system was replaced with alpha-numerical scheme. The room was acquired for many items and allows for almost doubling them compared to the previous revision. Expansion resulted from a single-letter or a group of letters marking, enabling creation of one hundred three-digit items per each disease (Picture No. 1). Of total 26 available letters, 25 were used and the letter U remained vacant for purpose of future addenda. This action was necessary also taking in account occurrence of new pathological conditions.

Chapter	**VIII**	**IX Body circulation system diseases**				**X**	**XI**
Class		I10 Essential (primary) hypertension	I11 Heart hypertension		I12 Renal hypertension		
Item			I11.0 Heart hypertension with congestion heart failure Hypertension heart failure	I11.9 Heart hypertension with congestion heart failure Hypertension heart failure NS			
Sub-item							

Picture No. 1 Example of coding system in ICD 10

In the countries that adopted the 10th revision too late, for example in USA, it was impossible to monitor and respond to threatening new pandemic viruses and bio-terrorism, e.g. SARS, Kreutzfeld-Jacobs disease, avian or H1N1 flu.[10] In this way, the latest revision has contributed to the development of the patient safety policies, providing detail view on the complications and death causes. It facilitates identification of medicine errors and method of their prevention; supporting and speeding up the innovation transfer between clinical research and practice. Last but not least, it contributes to the development of public health policies mainly through more detail knowledge on the pathological condition occurrence causes and to setting forth program priorities. In the terms of health insurance sector, it brings detail identification of cost, easier identification of potential areas of savings and high-quality service. Possibility to monitor the outputs, cost and performance of the provided healthcare services is important for continuous improvement in these areas.

Chapter No.	Code	Name
I.	A00-B99	Infectious and parasite diseases
Ii.	C00-D48	Tumors
Iii.	D50-D89	Blood and haemoplastic organs´ diseases and some disorders of immune mechanisms
IV.	E00-E90	Diseases of endocrine glands, metabolic and nutrition diseases
V.	F00-F99	Psychical and behavioral disorders
VI.	G00-G99	Nervous system diseases
VIi.	H00-H59	Diseases of eye and its adnexes
VIii.	H60-H95	Diseases of ear and its mastoid
IX.	I00-I99	Circulation system diseases
X.	J00-J99	Respiratory system diseases
XI.	K00-K93	Digestion system diseases
XIi.	L00-L99	Cutaneous and subcutaneous diseases
XIii.	M00-M99	Muscular, skeleton and connective tissue diseases
XIV.	N00-N99	Urogenital system diseases
XV.	O00-O99	Pregnancy, delivery and puerperium
XVI.	P00-P96	Diseases occurring during prenatal period
XVIi.	Q00-Q99	Genetic disorders, deformations, chromosomal aberrations
XVIii.	R00-R99	Subjective and objective symptoms and abnormal clinical and laboratory findings not otherwise classified
XIX.	S00-T98	Injuries, intoxications and other effects of external causes
XX.	V01-Y98	External causes of pathologies and mortality
XXI.	Z00-Z99	Factors affecting health condition and contact with healthcare services

Picture No. 2 Chapters of International Classification of Diseases – ICD 10.
Source: national Center of Healthcare Information, 2013

The dual classification scheme of causes and symptoms of diseases known as a cross (†) and star (*) system was implemented in the 9th revision but later replaced with optional use of the above marks in 82 homogenous 3-digit items. Such method enables capturing diagnostic terms with two codes, containing information about basic pathological process and symptoms/ complications in particular body organ, which allows for meaningful arrangement or structure.

The news introduced in the latest version included diseases that occurred as a consequence of a medical intervention. For example some diseases of endocrine glands and other metabolic disorders were included here.

E89 POST-INTERVENTION ENDOCRINE AND METABOLIC
DISORDERS NOT CLASSIFIE OTHERWISE

E89.0 Post-intervention hypothyreosis
hypothyreosis after radiation
post-surgery hypothyreosis
e89.1 Post-intervention hypo-insulinemia
Hyperglycemia after pancreasectomy
Post-surgery hypo-insulinemia
e89.2 Post-intervention hypoparathyreosis
tetania caused by missing
parathyroid glands
e89.3 Post-intervention hypopituitarism
Hypopituitarism after radiation
e89.4 Post-intervention ovarian failure
e89.5 Post-intervention testes hypo-function
e89.6 Post-intervention adrenal- cortical (-medular) hypo-function
e89.8 Other post-intervention endocrine and metabolic disorders
e89.9 Non-specific post-intervention endocrine and metabolic disorder

---

Picture No. 3 Example of post-intervention disorder coding

Creation of post-intervention disorder items at the end of particular chapters represented a significant innovation (Picture No. 3). These important conditions delimiting healthcare subject matter in its real meaning create a part of certain diseases of endocrine glands and metabolic disorders resulting from surgical removal of a body organ or its part, evtl. other specific diseases, for example post-resection syndrome following stomach resection. Post-intervention diseases that are non-specific and resulted as immediate complications after the intervention, for example pneumothorax and post-surgery shock, are classified in the chapter XIX. Injuries, intoxications and other effects of external causes.

Along with ICD, WHO developed other classification systems in order to enable international and national comparisons. Internationally adopted classifications enable storage, recalling, analysis and interpretation of data and their comparison within the population within time passing.

They allow for comparison between populations in the same time and compilation of internationally consistent data. [11] The following is a brief description of particular classifications description (Picture No. 4).

Related classifications	Reference classifications	Derived classifications
International classification of primary care (ICpc) [12]	International classification of diseases (ICD)	International classification of cancer diseases (ICD-O-3) **
International classification of external injury causes (ICeci) [13]	International classification of body functions, disabilities and health (ICF)[15]	ICD-10 classification of psychical and behavioral disorders***
Anatomic, therapeutic and chemical (Atc) classification system with defined daily doses [14]	International classification of health interventions (ICHI) in preparation	ICD application in dental medicine and stomatology (ICD-Da) [16]
ISO 9999 technical tools for disabled persons*		ICD application in neurology (ICD-10-NA)[17] Version ICF for children and youth (ICF-CY)[18]

Picture No. 4 Family of international classifications prepared by WHO

* ISO 9999:2007. Assistive products for persons with disability – Classification and terminology. It is available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=38894, ** http://www.who.int/classifi-cations/icd/adaptations/oncology/en/, *** http://www.who.int/substance_abuse/terminology/icd_10/en/ index.html

International classification of primary care (ICPC). [12] The 2nd version known under abbreviation ICpc-2 classifies data about patients and clinical interventions in the domains, considering the reason of medical appointment, discussed problems or diagnosis, interventions, and enables structuring the information according to particular episodes. It consists of 17 chapters, each divided in seven sections: symptoms and complaints, diagnostic, screening and preventive procedures, therapy, medications and procedures, analysis results, administration, recommendations and diseases. It is an outcome of long-term work of the team WONCA.[9]

International classification of external injury causes (ICECI) [13] allows for systematic description of injury causes. Originally, it was intended for injury prevention

[9] http://www.globalfamilydoctor.com/wicc/about.html

based on statistical data obtained for example through surveillance of patients at the traumatology outpatient´s ward. Coding structure allows for capturing the cases independently from each other. Therefore, it is possible to allocate a code to objects and substances that caused injury regardless other items coding, for example purposefulness. Particular modules are thematically structured. Main module is suitable for surveillance; module *sport* is focused on sport-related injuries.

Anatomic, therapeutic and chemical (ATC) classification system with defined daily doses [14] represents a tool for improved use of pharmaceuticals based on knowledge of the situation in the population. It has proved useful in the studies that compared the use of pharmaceuticals on national and international level. It was helpful at setting forth long-term trends, as well as estimation of the effects of particular policies. Its value was confirmed during monitoring of safety of pharmaceuticals..

ISO 9999 technical tools for disabled people. This international standard classified technical tools for disabled people, focusing on the tools intended for individuals. The classification consists of three hierarchy levels. Each class, sub-class or ward has a code and name allocated, attached with explanation text or indicator/ reference to other classification part, if necessary.

International classification of body functions, disablements and health (ICF) [15] classifies health domains in the terms of human body, individual and the society. It distinguishes between two basic lists; the first one referring to the list of functions and structures and the second one containing the activities and participations. Moreover, it contains the list of surrounding factors and allows for measurement of health and disablement of an individual or population. ICF considers that everybody can face health disorders and acquire certain degree of disablement. Moreover, it takes in account also social aspects of disablement instead of treating them only as medical or biological entities. Last but not least, it covers also environmental factors and their effect on the human life.

International classification of health interventions (ICHI)[10] has been currently prepared. It will be aimed at providing WHO member countries, healthcare providers, managers and scientists with a common tool for observation of the structure and development of interventions in the area of health. Originally, it was limited to surgical procedures only, and published under the name *International classification of medical procedures (ICPM)*. Further development continued in Germany under the name *Operationen – und Proze-durenschlüssel OPS*.[11] [19]

International classification of cancer diseases (ICD-O-3)[12] is mainly intended for registration of oncology patients and enables coding of tumor

---

[10] http://www.who.int/classifications/ichi/en/
[11] http://www.dimdi.de/static/en/klassi/prozeduren/ops301/index.htm
[12] http://www.who.int/classifications/icd/adaptations/oncology/en/

location and its histology. It is used to multi-layer classification of the occurrence site, histology and behavior – the nature and degree of neoplasm. The topography uses ICD 10 and adds it with details of localization mainly in case of haemoplastic and reticulo-endotelial tumors. Morphology describes tumor histology with 5-digit code, thereof the first 4 digits are specific for histological term. The fifth digit describes histological differentiations. Classification is often combined with ICD, mainly when used in epidemiology. [20]

ICD10 Classification of psychic and behavioral disorders.[13] Chapter No. 5 of ICD 10 contains detail classification of more than 300 psychic and behavioral disorders. There are two versions available, Clinical Descriptions and Diagnostic Guidelines, and Diagnostic Criterions of Research. The first one describes major signs and symptoms of each disorder, along with other, side signs. It also contains the diagnostic guidelines. The second version serves as a supporting tool for research based on the criterions enabling selection of individuals with similar symptoms and other characteristics.

Application of ICD in stomatology and dental medicine (ICD-DA)[14] [16] represents a system of data coding and classification related to buccal cavity and teeth defects. It is deduced from ICD10 and recommended for use together with ICD10. Relevant items from ICD 10 were extended by the fifth digit that makes the finding more specific.

ICD application to neurology (ICD-10-NA)[15] [17] complements ICD-10 and makes it more precise through added further 3 signs to the codes. The system is attached with detail list and definitions.

Version ICF for children and youth (ICF- CY)[16] [18] is an ICF version modified for children and youth, enabling to code body functions and structures, restrictions to activities that occur during childhood, youth and adolescence.

Clinical version of ICD-10-CM

Despite of significant progress in the form of issued 10th revision of ICD, it is still insufficient for clinical use. The revision amendment was initiated at the national Center for Health Statistics (USA) where the 10th revision was only lately implemented, for addition of the clinical measures under the name ICD-10-CM. This amendment is complementary to the original revision so as the diseases can be classified, and it makes diagnostic categories more specific. The system contains more than 68 thsd. codes. It is presumed that ICD-10-CM will be used for healthcare quality and outputs

[13] http://www.who.int/substance_abuse/terminology/icd_10/en/index.html

[14] http://apps.who.int/bookorders/anglais/detart1.jsp?sesslan=1&codlan=1&codcol=15&codcch=3086

[15] http://apps.who.int/bookorders/anglais/detart1.jsp?sesslan=1&codlan=1&codcol=15&codcch=2276

[16] http://apps.who.int/bookorders/WHP/detart1.jsp?sesslan=1&codlan=1&codcol=15&codcch=716

evaluation. [21]

The following revision

WHO has initiated the 11th revision of ICD. [22] The new version should reflect the progress in medicine and all health-related sciences. As well, it should emphasize the usefulness for clinical purposes and healthcare systems in the terms of easier use, support of clinical decisions and management that would further simplify its integration in the routine practice on all levels from the primary care up to research. Last but not least, it should be useful in the area of source deployment strategy and monitoring of results based on recorded death rate, disease rate and other health parameters. As well, it should be compatible with other classification schemes and information systems (IS). Final version of the 11th revision should be submitted for approval to WHO General assembly in year 2014.

Systemized Medicine Nomenclature – clinical terms SNOMED-CT

SNOMED CT is considered the most complex, multi-lingual terminology of the clinical medicine and healthcare. It was developed for application in the systems of electronic health records to provide for consistent, reliable and comprehensive record of clinically relevant data. [23]

Start

SNOMED was launched in 1965 as SNOP (Systemized Nomenclature of Pathology) and later it had expanded to other areas of medicine. It was developed jointly in the Great Britain and USA. In 2007, *International Organization for Development of Healthcare Terminology Standards* IHTsdo overtook the SNOMED copyrights,[17] and has been developing it to date. SNOMED Clinical Terms® (SnoMED CT®) has been periodically updated on basis of requirements submitted by users from the whole world. Two adjustments are made per a year.

Structure

From abscess to zygote, the extent of concepts covers more than 300 thsd. independent units. The term *concept*[18] shall be interpreted as a term

[17] http://www.ihtsdo.org/

used in medicine or healthcare sector and associated with codes, while a single code can cover more synonymic terms. [24] Currently, there are approx. 800 thsd. concepts. The terms in this hierarchy refer to the results of clinical observation, evaluation or assessment and can contain normal and pathological clinical conditions. An example of a term included in the clinical findings is pure sputum (finding), normal breathing sounds (finding) or wrong posture (finding).

Codes are created pursuant to the principle "one code one meaning" and vice versa. It consists of a chain of 6 – 18 signs. For example, 22298006 means „*Heart Attack – myocardial ischemia (MI)*" and 399211009 means „MI in the past". The terms are in humans´ heads and the codes represent terminology, however both terms and terminology describe the real world. [25] Description is a word allocated to the term. There can be two descriptions with the same text, for example immuno-suppression → immuno-suppressive treatment (procedure), description ID = 507152014 or imuno-suppression → imunom-suppression (finding), description ID = 63394015.

Relations can be also of various types. Definition relations must be a true statement of a term. Qualificators are added to specify the term. Historical data provide for indicator of the obsolete and currently used term, enabling to add information that has no direct relation to the definition. Moreover, the system contains more than a million of relations between the units, entered as: *Diabetes Mellitus IS_A glucosis intolerance disorder*. Operator *IS_A* is an expression of certain relation or allocation. In this way, the whole hierarchy of terms and correlations can be built up. The extent of the coding system can be imagined as follows: If a person spends 1 minute with studying each description and worked 8 hours a day, 5 days a week, it would take him approx. 305 weeks, i.e. approx. 6 years to read all descriptions.

It can be recognized from this very brief introduction that such a built system is dynamical contrary to ICD and thus suitable for use in day-to-day clinical praxis.

Slovakia

SNOMED-Ct use has been considered in Slovakia within the development of a patient´s e-record included in the project e-Health. [26] Slovak authors made efforts to use SMONED already at the end of the 70s in the 20th century for purposes of medical information processing via computer. SNOMED version was translated in Slovak language at the Research Institute of Medical Bionics in Bratislava and PC program was

[18] ISO-1087 (1990) Term (concept): thought unit created through abstraction on basis of common properties of a group of objects. ISO-1087 (2000) Term (concept): cognition unit created through unique combination of typical properties. Typical property as such is defined as: abstraction of property of the object/ group of objects.

created in PC environment MUMPS as a coding support. [27-30]

Payment for diagnosis (so called DRG)

Description of classification and terminology systems presented in the previous sections stressed out that they could be used also for increase and monitoring of healthcare quality and as a support of medicine-related decision making. In the last section, we will describe the system that was created originally with the aim to reduce large amount of particular items in the patients´ records, mainly diagnoses, to limited amount of the groups with certain common properties. [31, 32] It had gradually evolved to a language that allows for joining financial and clinical aspects of the care of hospitalized patient. This need has been emphasized with the fact that average hospitalization cost pursuant to OECD[19] represents approx. one third of healthcare cost in OECD European countries in 2008, within 26% in Slovakia and 47% in Sweden. [33] In USA, so called "fee-for-service" has been traditionally paid for hospital medicare. This led to charging unacceptably high prices for sometimes useless and inefficient procedures. DRG implementation resulted in significant savings and represented transition from retrospective to prospective payments. On the contrary, services were paid in Europe from the budget. What caused DRG spread also in Europe? It was as well the price of services that was, however, based on exact calculation of the patient´s cost, taking in account measurable characteristics of the patient. Similarly, they can be applied also to assessment of provided services quality and efficiency. [34]

DRG structure

The concept *case-mix* is the basis of DRG philosophy. *Case* also refers to a patient; mix has its usual meaning. Expression *case-mix* used to be interpreted as a structure or arrangement of hospitalized patients/ cases (sometimes "mix of patients") or the sum of weight ratio of all cases concluded in certain time period (e.g. one year) in the hospital or other healthcare facility. In the 2nd case when the weights are summed up and divided by the number of cases treated during certain period of time, the result refers to the *Case-Mix Index (CMI)*. Thus, case-mix refers to classification of cases in the disease-based groups according to preset criterions. DRG is one of the group of sorting analytical case-mix systems enabling description and sorting of hospitalization cases according to diagnoses and selected performed interventions

19 Organization for Economic Cooperation and Development

in the groups with similar clinical progress and similar economic cost. Patient´s diagnosis or medical intervention is a main criterion for inclusion in DRG groups. Inclusion of more cases in one group is made on basis of similar therapeutic method and similar therapy cost. Weight is then allocated to each group, expressing clinical complicatedness of the case and characterizing the rate of usual financial sources requirement (cost) for healthcare provided to the patients within particular DRG group. This weight allocation is usually made on basis of the data obtained during preceding period.

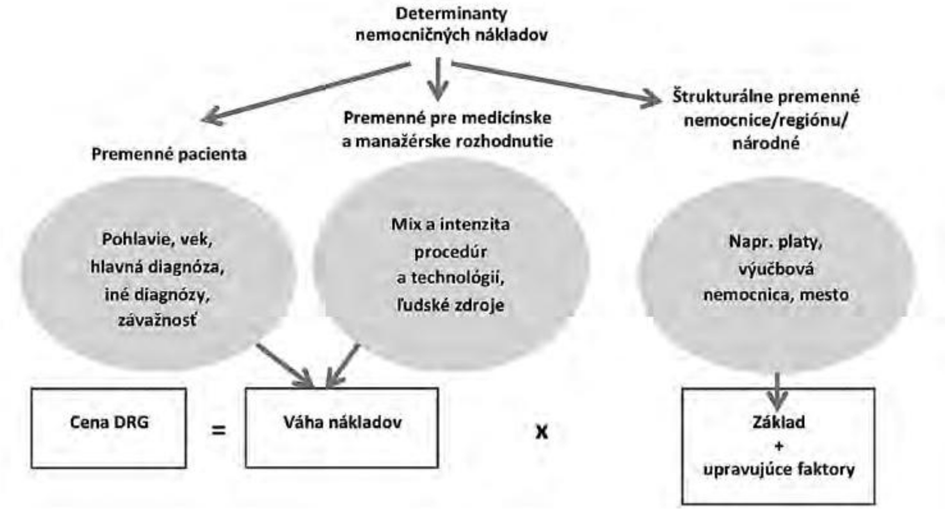
Patient´s classification represents the first step within the DRG specification (Picture No. 5), performed mainly on ICD basis. Then selected data should be collected from actual patient´s record for certain period. Price specification represents more complicated procedure with applied statistical procedures, mainly regression analysis (Picture No. 6).

Picture No. 5 DRG specification process (pursuant to [34])

DRG introduction is a major step towards making the medical outputs price, healthcare provision and evaluation of patient´s care quality more transparent. Establishment of uniform price for a group of outputs naturally hinders efforts for wasting medical procedures, since these are not included/ not paid in the DRG price. However, it is not an all-purpose tool and its application doesn´t automatically solve all healthcare system problems. The task of top management, leadership and personal efforts to assume responsibility for own performance remain the most important tasks.

Coding quality

Coding itself is not a goal but the tool of statistical observation support. Its quality depends on the system proposal but mainly on the input data reliability.



Picture No. 6 Calculation of DRG values (pursuant to [34])

The slogan „Garbage in and garbage out",[20] still counts, thus if there is coded nonsense or missing data, it results again in incorrect conclusion. Error can be random, like in case of all operations with large amount of data. In case of death causes coding, it is often systematic, for example resulting from various approaches and interpretations of code application.

Rothenberg and Aubert [35] investigated the effect of transition from the 8th to 10th ICD revision in case of mortality associated with ischemic heart disease and high blood pressure. We know that a physician concluding a death cause often chooses the first one in order despite of knowing that long-term developing hypertension was actual death cause. The authors found out that implementation of the new revision change the coding method, resulting in changed ratio of two stated pathologies. Quality of collected data and their coding then significantly affects possibilities of their use for example for health strategy support [36]. In the stated ischemic heart disease and hypertension example, relative hypertension drop mentioned in the death causes could result in faulty conclusion on efficient anti-hypertension campaign and to transfer the funds intended for such campaign to other entity. Such decision could seriously affect the search for people suffering hypertension and their therapy. [37] There are indications of insufficient coding quality also in Slovakia. [38] Many countries make efforts to define real death causes, implementing measures aimed at increasing the quality of death causes recording. [39 – 42]

20 GIGO means "garbage in, garbage out". The phrase has become worldwide known as a warning that IT, if we feed it with incorrect data, is not able to recognize what it is about and processes also garbage, giving incorrect results.

Summary

The reader was acquainted with the diseases terminology and development of diseases classification. We presented the goals of these programs as well as main systems currently applied in practice. Major part of the chapter was dedicated to classification of diseases as a main tool of identification of health condition in particular territorial wholes. Its application to disease sub-classes, e.g. cancer, psychiatric and other ones, was indicated. As well, the authors pointed out the classification development in the clinical direction and its application during healthcare purchase.

References

1. WHO. *History of the development of the ICD*. [cited 2010 28.12.]; Available from: http://www.who.int/classifications/icd/en/HistoryOfMKCH.pdf.
2. Israel RA. *The International Classification of Disease. Two hundred years of development.* Public Health Rep, 1978. **93**(2): p. 150 – 152.
3. Blaxter, M. *Health and Lifestyles*. 1st ed. 1990, London: Routledge.
4. Young, T. K. *Population Health: Concepts and Methods*. 2nd ed. 2004: Oxford University Press, USa.
5. Bujalkova, M. *Eponyms in medical terminology.* Bratisl Lek Listy, 1999. **100**(5): p. 271 –272.
6. *International classification of diseases, injuries and death causes.*1948, Prague: Ministry of Healthcare, Orbis Praha.
7. *Report of the Commission of Revision of the Bertillon Classification of Causes of Death.* Public Health Pap Rep, 1899. **25**: p. 353 – 356.
8. Greenberg SJ. *The "Dreadful Visitation": public health and public awareness in seventeenth-century London.* Bull Med Libr Assoc, 1997. **85**(4): p. 391 – 401.
9. WHO. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Version for 2007*. 2007 [cited 2010 30.12.]; Available from: http://apps.who.int/classifications/ apps/icd/icd10online/.
10. Schwend G. *Expanding the code. The methodical switch from ICD-9-CM to ICD-10-CM will bring both challenges and rewards to healthcare.* Health Manag Technol, 2007. **28**(6): p. 12, 14.
11. Madden, R., Sykes, C., Ustun, T. B. *Family of International Classifications: definition, scope and purpose*. 2007
12. Hofmans-Okkes, I. M., Lamberts H. *The International Classification of Primary Care (ICPC): new applications in research and computer-based patient records in family practice.* Fam Pract, 1996. **13**(3): p. 294 – 302.
13. ICECI Coordination and Maintenance Group. *International Classification of External Causes of Injuries (ICECI). Version 1.2*. 2004; Available from: http://www.rivm.nl/who-fic/ICECI/ICECI_1-2_2004July.pdf.
14. WHO. *Introduction to Drug Utilization Research*, W.I.W.G.f.D.S. Methodology, W.C.C.f.D.S. Methodology, and W.C.C.f.D.U.R.a.C.P. Services., Editors. 2003: Oslo.
15. Stucki G et al. *Application of the International Classification of Functioning, Disability and Health (ICF) in clinical practice.* Disabil Rehabil, 2002. **24**(5): p. 281 – 282.
16. Bezroukov, V. *The application of the International Classification of Diseases to dentistry and stomatology.* Community Dent Oral Epidemiol, 1979. **7**(1): p. 21 – 24.

17. van Drimmelen-Krabbe , J. J. et al. *The application of the International Statistical Classification of Diseases to neurology: ICD-10 NA*. J Neurol Sci, 1998. **161**(1): p. 2 – 9.
18. Lollar, D. J., Simeonsson, R. J. *Diagnosis to function: classification for children and youths*. J Dev Behav Pediatr, 2005. **26**(4): p. 323 – 330.
19. Stausberg, J. et al. *Classifications in routine use: lessons from ICD-9 and ICPM in surgical practice*. J Am Med Inform Assoc, 2001. **8**(1): p. 92 – 100.
20. Burgun, A., Bodenreider, O. *Issues in integrating epidemiology and research information in oncology: experience with ICD-O3 and the NCI Thesaurus*. AMIA Annu Symp Proc, 2007: p. 85 – 89.
21. Barta A et al. *ICD-10-CM primer*. J AHIMA, 2008. **79**(5): p. 64 – 66; quiz 67 – 68.
22. WHOFIC and Network. *Production of ICD-11: The overall revision process*. [cited 2010 4.1.]; Available from: http://www.who.int/classifications/icd/MKCHRevision.pdf.
23. Dolin, R. H. *SNOMED at the point of service: the challenges and opportunities of SNOMED-encoded data*. J AHIMA, 2010. **81**(3): p. 42 – 43.
24. IHTSDO. *SNOMED CT Clinical finding*. [cited 2011 5.1.]; Available from: http://www.ihtsdo.org/snomed-ct/snomed-ct0/snomed-ct-hierarchies/clinical findingdisorder/#c1509.
25. Spackman, K. *SNOMED Clinical Terms Fundamentals*. 2007, IHTSDO.
26. Šagát, T. et al. *ELEKTRONICKÝ CHOROBOPIS*. 2006 [cited 2011 5.1.]; Projekt: e – Health: [Available from: http://www.nczisk.sk/buxus/docs/eHealth/studie/STUDIA_EHR.pdf.
27. Rusnak, J., Kristof, J. *Clinicians diagnostic and treatment activities algoritmization (author's transl)]*. Bratisl Lek Listy, 1978. **69**(2): p. 124 – 129.
28. Ciganek, L. et al., *Computer-assisted recording of the objective neurologic finding*. Eur Neurol, 1977. **16**(1-6): p. 165 – 171.
29. Rusnak, J. E. *Medical information science and application: the convergence of modern technologies*. Med Inform (Lond), 1984. **9**(3-4): p. 215 – 216.
30. Rusnák, J., Rusnák, M. *Cybernetic methods*, in *Bio-physics for Physicians*, J. Holan, Editor. 1982, Publishing agency Osveta: Martin. p. 348 – 419.
31. Fetter, R. B. et al. *Case mix definition by diagnosis-related groups*. Med Care, 1980. **18**(2 Suppl): p. iii, 1 – 53.
32. Goldfield, N. *The evolution of diagnosis-related groups (DRGs): from its beginnings in case-mix and resource use theory, to its implementation for payment and now for its current utilization for quality within and outside the hospital*. Qual Manag Health Care, 2010. **19**(1): p. 3 – 16.
33. Cylus, J., Irwin, R. (2010). *The challenges of hospital payment systems*. EuroObserver, **12**, 3, p.1 – 3.
34. Scheller-Kreinsen, D., Geissler, A., Busse, R. (2009) *The ABC of DRGs*. EuroObserver **11**, 4, p. 1-5
35. Rothenberg, R. B., Aubert, R. E. *Ischemic heart disease and hypertension: effect of disease coding on epidemiologic assessment*. Public Health Rep, 1990. **105**(1): p. 47 – 52.
36. Sirken, M. G. et al. *The quality of cause-of-death statistics*. Am J Public Health, 1987. **77**(2): p. 137 – 139.
37. Rosenberg, H. M. *Cause of death as a contemporary problem*. J Hist Med Allied Sci, 1999. **54**(2): p. 133 – 153.
38. Baráková, A., et al., *Objectification of death causes and statistical data relevance in the obduction records. Information-analytic study*, 2011, national Center of Healthcare Information: Bratislava.
39. Harteloh, P., de Bruin, K., Kardaun, J. *The reliability of cause-of-death coding in The Netherlands*. Eur J Epidemiol, 2010. **25**(8): p. 531 – 538.

40. Franca, E. et al. *Evaluation of cause-of-death statistics for Brazil, 2002-2004.* Int J Epidemiol, 2008. **37**(4): p. 891 – 901.
41. Chen, L., Walker, S., and Tong, S. *The impact of the variation in death certification and coding practices on trends in mortality from ischemic heart disease.* Aust Health Rev, 2002. **25**(4): p. 189 – 197.
42. Lu, T. H., Lee, M. C., and Chou, M. C. *Accuracy of cause-of-death coding in Taiwan: types of miscoding and effects on mortality statistics.* Int J Epidemiol, 2000. **29**(2): p. 336 – 343.

Chapter 4

Data storage and DB systems

Chapter objectives

Data administration belongs to the skills that an ambitious public healthcare specialist should avoid. We are usually satisfied with services provided by any of the programs MS Office, be it Excel or Word. Working with larger data volume, we often encounter restrictions in size or complicated structure, or impossible searching. DB systems enable data storage, handling, using the tools for data sorting out, defining the selection terms or proposing the data storage and retrieval forms.

Before starting to deal with DB work tools, we will present the overview of the informatics related to the data. Reader could complain on the time wasting because of the subject matter that is taught at elementary or high school. However, repetition has still been the basis of study.

Data

We already explained the difference between the data and information. Anyway, we didn´t mention the tools that facilitate work with data and their transformation to information. We are going to skip the pre-computer era, nevertheless majority of data had been transformed before with a paper and pencil exclusively (the data were recorded and transformed even before the paper discovery). Let´s remind Mikuláš Kopernik (1473 – 1543) who placed the Sun in the center of the Universe (About Heavenly Spheres Movement). We should think about Johannes Kepler as well (1571 – 1630), who understood and deduced the principles of the Earth circulation around the Sun from observations and recording movements of the stars in the tables with the digits (also using the tables left by Tycho de Brahe (1546 –1601).

Isaac Newton (1643 – 1727) referred to them when deducing the gravitation laws. John Snow (1813 – 1858), founder of epidemiology who plotted the deaths in the central London map, found out where the cholera epidemic stemmed from. [1] Neither of them had PC or calculator available but they make all calculations in their heads and wrote them down.

Today it is almost impossible to make such discoveries without computer. We cannot imagine that ledgers were kept manually in the past, columns and rows were summed up and the results were transferred from one to another page. The first calculations significantly facilitated the work of accountants, engineers and scientists. Today it seems impossible that such major discoveries were made without computers. Let´s ask what discoveries or procedures are impossible without computer technique. In the area of health and medical sciences, it is definitely computer tomography and related displaying methods; or processing large files in the area of public health. For example, computer allowed for demonstration of relationship between smoking and lung cancer on the specimen of more than 10,000 people. [2] Before it was impossible to process such large data files without error that completely devaluated the result. Spread of statistic methods; mainly the causality study using logistic regression and further multi-variant methods would also be impossible without PC use.

How could a computer do it? The basis of PC operation is its ability to work with information in the form of digits 1 (ON) and 0 (OFF). Computer is able to work nonstop, fast and contains billions of such switches. Today it is by far no miracle; we have computer in our cellphone, in the car, at home, at work or during fun. Speaking of data in computer; the switch calls "bit" being switched ON or OFF. To be able to work with bits using letters or digits, eight bits create one "byte". Since we get 256 variations from eight combinations of bits (two with exponent 8 or $2^8 = 256$), we can allocate a letter, digit or sign to each of them. There are more allocation methods but the system ASCII has been used most frequently, present in majority of contemporary computers, tablets and i-phones. Since originally it is a US table, it didn´t contain alphabetical signs with diacritic marks. ASCII was spread and currently covers majority of the world alphabetical systems.

Programs and data

Computer is not able to work without operating manual, however. It is called program and it provides the hardware with detail instructions of what to do. The operation is performed with data presenting reality, be it photos, music, words, sentences or digits. The reality is of a form of data and it is the subject of our interest. Obtained data must be stored in computer so as they could be retrieved at any time.

To work properly, computer needs power supply. When you switch computer off, all have done so far would be lost if you didn´t use any of the data storage methods. Many people have bad experience with power outage and the result of efforts made.

Two processes are required for data storage and use. The first one is called *data entry* and the second one – *data reading.* The first process enters/ records the data on the medium with signs 0 and 1 and the second process read them. Physical environment of data entering and reading is called *medium* (storage medium). Paper strips were used as a medium in the past and the data were punched into them and stored on paper cards where the data were recorded through punched holes. These cards were later replaced with magnetic strips and magnetic discs. Nowadays, two medium types are used most frequently – magnetic and optical. Magnetic media are easily rewritten but are prone to errors and defects and vibration sensitive. They are used in computers and represent most frequently used media. Well known CD discs work with the laser, thus called optical discs. They perfectly store data but rewriting is not so optimal. They are mostly used for creation of backups for long-term storage; gradually extending to other media, for example SSD (Solid State) discs without mechanics (popular USB keys).

Programs for work with data – DB systems

Data quantity has been increasing so fast that it´s hard to keep overview of it. Therefore, programs were development for working with data that we call database (DB) systems. They are environments that significantly simplify work with data, i.e. creation, storage and retrieval of information from DBs. DB system called also data bank is a program system serving for effective storage, modification and retrieval of huge amount of persistent data. [3] DB is a set of information organized so that data are easily available, controlled and updated.[4] DBs used to be classified according to content type: bibliographical, full-text, numerical or with graphical content. Sometimes they are classified according to method of data access. Relation DB is mostly used and it is based on the system of mutually interconnected tables. Relation DB contains data defined so that they can be reorganized and accessible in various ways. Distributed DB consists of many nodes with scattered data or replicated among various network points. Object-oriented DB works with data defined in the form of objects in the terms of classes and sub-classes. [5] DBs usually contain data records or files aggregation; in our case they can be files containing data of patients at particular hospital wards or overview of physicians´ outputs at hospital wards, or total reported infectious disease that occurred in the region.

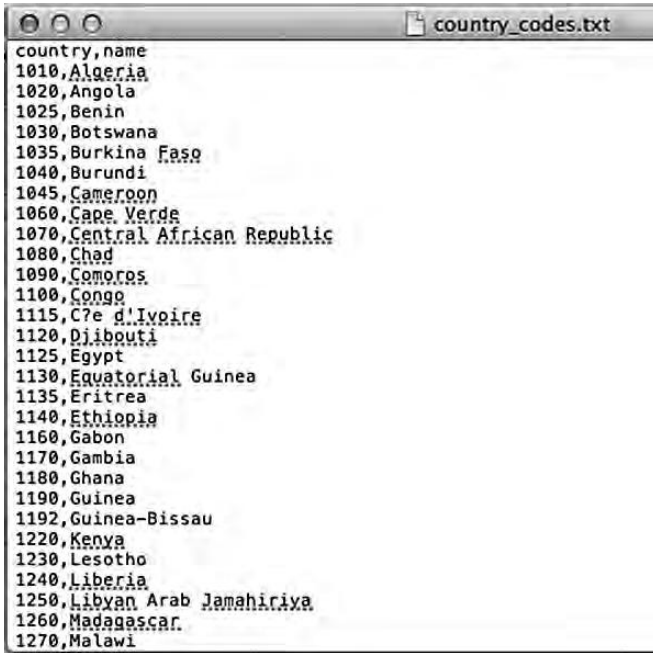
SQL

Data represent such an important part of modern programming that there are programming languages intended for work with DB. *Structured Query Language (SQL)* is a primary standard for DB languages.

SQL is standardized program for DB creation, storage of information in DBs and retrieval of information from DBs. Program developers usually create data structure in SQL at first and then write the program in another language in order to obtain access to the data. As soon as you learn SQL, you can apply it easily to almost all programming languages. Basic SQL terms remain the same regardless the DB type you work with. Majority of SQL orders can be used without any modification in the applications Microsoft Access, Microsoft SQL Server, Oracle and many others packages of related DB systems (RdBMS). In this section, we will use DB system FileMaker, which is commercially available for Windows and MAC platforms. However, reader can use any other, commercial or other freely accessible package, for example LibreOffice. [6]

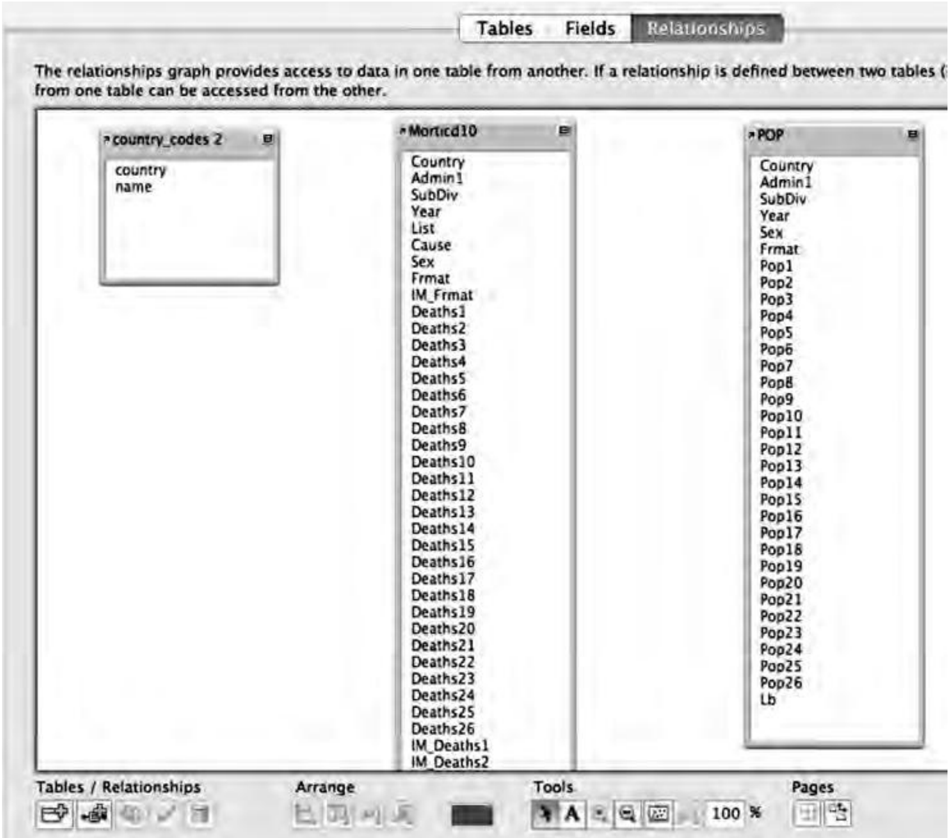
Simple DB

Most DB programs allows for use with two DB types at once: simple and elation ones. Simple DB consists of one or more tables, each of them containing one or more fields. Each table is completely independent from all remaining tables within the DB. For example, we have mortality DB where every table contains data from other country or each row in DB has an independent table. We will download the file *Country codes*, packed In ZIP format, on basis of mortality DB from WHO - *WHOSIS* [7]. After unpacking


country_codes.txt

```
country,name
1010,Algeria
1020,Angola
1025,Benin
1030,Botswana
1035,Burkina Faso
1040,Burundi
1045,Cameroon
1060,Cape Verde
1070,Central African Republic
1080,Chad
1090,Comoros
1100,Congo
1115,C?e d'Ivoire
1120,Djibouti
1125,Egypt
1130,Equatorial Guinea
1135,Eritrea
1140,Ethiopia
1160,Gabon
1170,Gambia
1180,Ghana
1190,Guinea
1192,Guinea-Bissau
1220,Kenya
1230,Lesotho
1240,Liberia
1250,Libyan Arab Jamahiriya
1260,Madagascar
1270,Malawi
```

Picture No. 1 Codes and names of countries for WHO mortality DB [7]



Picture No. 2 DB structure with particular tables and field names

we get a file in whose name we have to add extension *.csv* . This contains the country code and name separated with comma; names of both columns are in the first row (Picture No. 1).

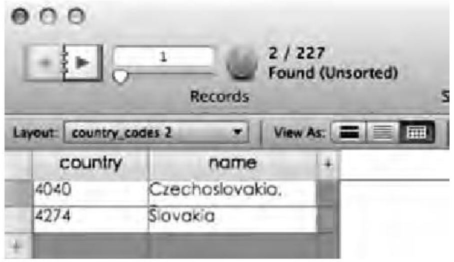
Afterwards, we insert the file content in DB. Process of data uploading in the DB file can slightly differ depending on the use of particular environment, thus we are not going to mention it here anymore. Who wants to test these options should use the manual applicable to the used program environment. We shall create the table of the age structure and read the specific mortality data based on the causes, age and gender. Resulting DB shall be called „summary“. While the file with codes is rather small, containing only 227 entries, the file with data related to population according to age structure and gender in particular years is much larger and contains 8475 entries. The largest one – death causes according to age structure, gender and causes in particular years has more than 2 million entries. In case of the first file, we can work with it in Excel or other table editor. The other file can be still used without DB tool; but the third one is impossible to open in Excel and it indicates error upon attempt for reading. In this case, DB tool use is the most suitable solution.

As soon as all three files are uploaded in DB, we can display them together with arrays names (Picture No. 2). Reader should notice that the array „Country“ is on the head of each table. It is ready for eventual interconnection of tables and creation of the relation DB from a simple – flat DB. This would significantly simplify the data selection for further processing. However, its preparation is more demanding and expects the developer´s experience in this area. Thus we are not going to deal with it here anymore.

When we have a DB ready, we can start using it. Please notice that WHO uses numerical codes for particular countries. The first step is to find the codes with applicable data for Slovakia. If we want to work with data from the period of Czechoslovakia, we will have to seek the codes for both Slovakia and Czechoslovakia.

Thus it´s not enough to search for the text „Slovakia“, but choose optional selection including other letter before or behind the name. We can do it through operator* saying: “search the chain “Slovakia” in the text” and allow further text before or behind it”. Thus the operator will look like: **“Slovakia“**.

country	name
1010	Algeria
1020	Angola
1025	Benin
1030	Botswana
1035	Burkina Faso
1040	Burundi
1045	Cameroon
1060	Cape Verde
1070	Central African
1080	Chad
1090	Comoros
1100	Congo
1115	C?e d'Ivoire
1120	Djibouti
1125	Egypt
1130	Equatorial Guinea

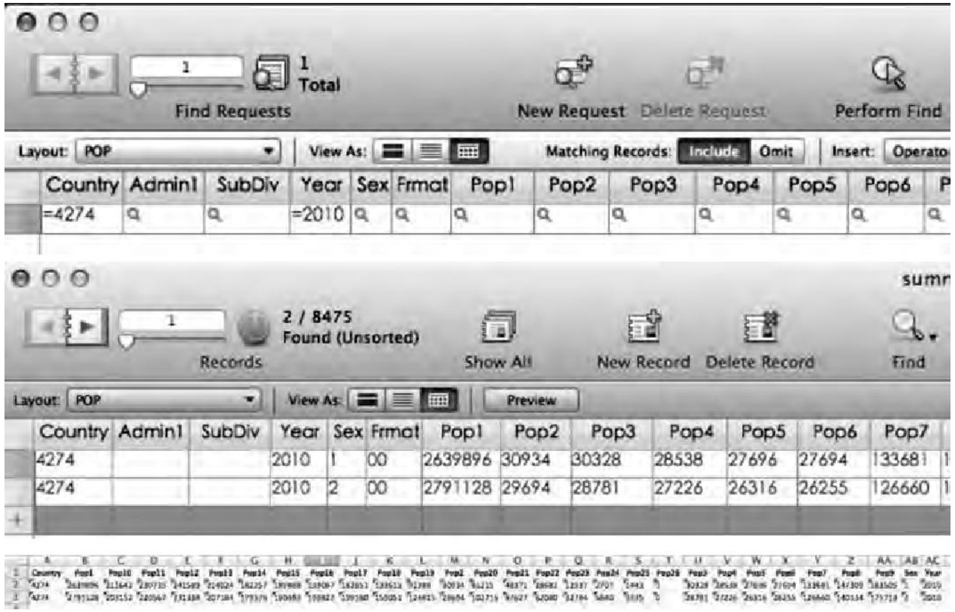


country	name
4040	Czechoslovakia.
4274	Slovakia

used operator
“Slovakia“

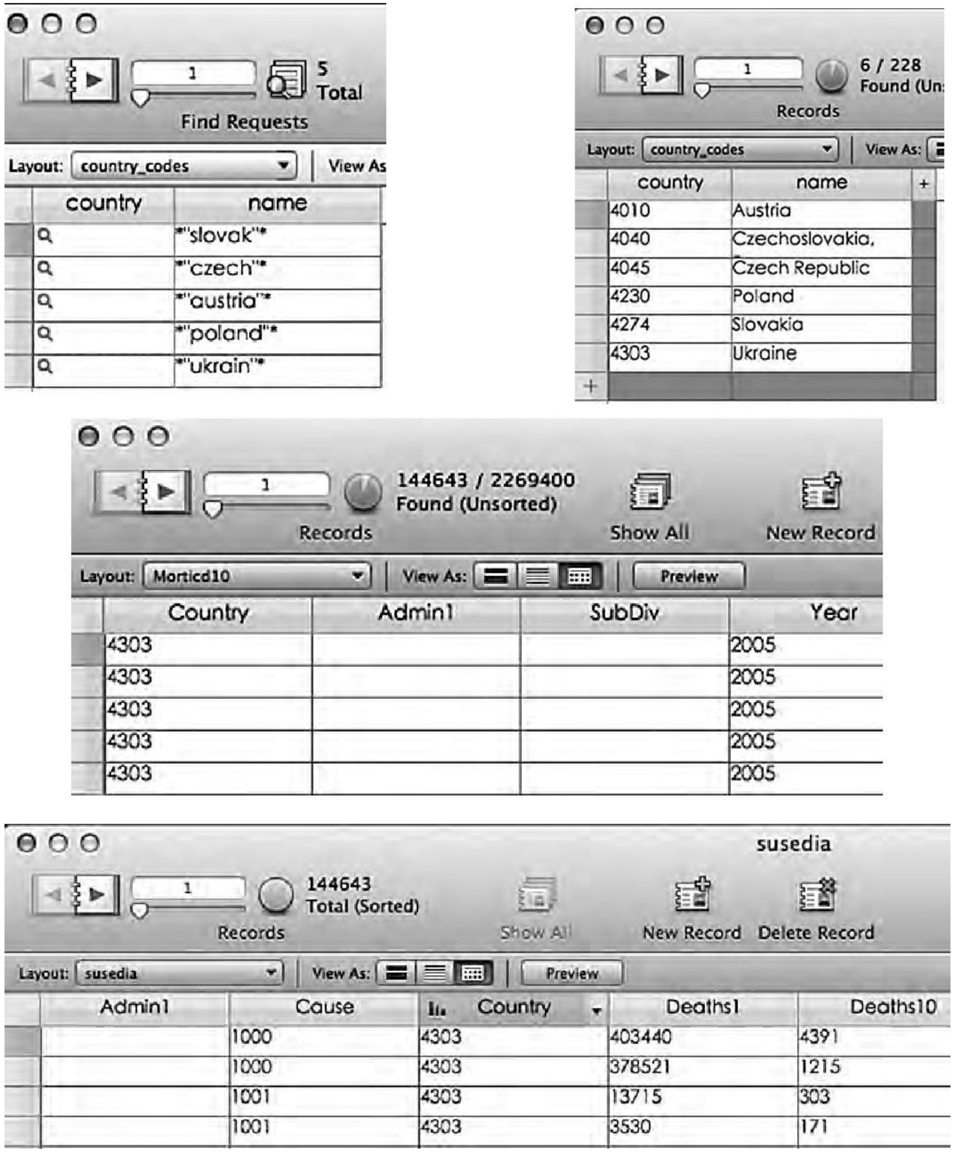
Picture No. 3 Country codes in WHO DB and search for Slovakia

The table with code for Slovakia and Czechoslovakia should be the result (Picture No. 3). Further we can work with the found codes that will allow for required data retrieval, together with those enlisted in the accompanying DB document (Picture No. 4). We can work with such retrieved data either directly in the DB environment, some systems allow for graphs drawing, or format message in the form of tables. If we need to work further with data, we can store them or export in other formats, for example in *.xlsx*, or *.csv*. We often need to retrieve only the data from the basic DB that we plan to work with, instead of dealing with entire huge DB. Thus we will create a derived DB based on the original one, where we will store exclusively the data we intend to work with. For example, we will choose the data of countries neighboring with the Slovak Republic from the huge DB of mortality containing more than 2 million entries (Picture No. 5).



Picture No. 4 Formulation of selection terms and result and export in Excel file

We will get much smaller file that is much easier to work with. The process is similar to the previous cases; we just have to apply slightly more complicated rules. At first, we enter the table with country codes and choose those corresponding to our requirements: Slovakia, Poland, Hungary, Austria, Czech Republic and Ukraine. Subsequently we choose from the death DB. Since we will work with the data later, we will keep data for all chosen years.



Picture No. 5 Two pictures in the top row document selection of codes for countries neighboring with Slovakia. Underneath, there is a selection result with number of entries reduced to 144 thsd. from more than 2 million. After saving in the form of a separate DB file, the number of entries remained unchanged.

Afterwards we export the selection in the new DB file, and name it for example "neighbors". We chose the codes for selected countries through stated parts of words creating the country name, and their alignment in the selection term.

Then we applied the found codes to the formulation of the terms of selection of data about deaths according to causes. Since the terms of other variables weren´t specified, we chose data for all years, diagnoses and both genders. We received more than 144 thsd. entries. We should take in account that there are still 2 million of entries in the file, despite of the selection. It will change as soon as we export the selected data in the new DB. Reader can ascertain himself about the huge difference between the size of original DB and the new, smaller one. This will of course influence the speed of working with DB.

We presented the methods of working with large data files through DB programs. Considering the situation progress when a clerk working in the public healthcare system has more and more data available, the authors recommend the readers making effort to obtain at least fundamental skills in the area of work with DB programs, as well as work with MS Word or MS Excel.

MySQL

MySQL is a system of work with DBs, intended mainly for simple DBs with large data amount. It´s advantage is in adjustment mainly to web applications. It contains orders of creation, cancellation and editing of a table; as well as orders of work with the data, e.g. addition, deleting, editing and search for data. SQL is the basis through which it communicates with DBs type MySQL, Microsoft SQL server, PostgreSQL, Oracle, etc. MySQL was created by the Swedish company MySQL AB [8] and it has acquired ever more popularity thanks to its stability, speed, simplicity and mainly because it´s free of charge for all platforms. It can be installed with Linux, MS Windows and other operating systems. MySQL is not demanding in the terms of PC sources. Relation model defines the stored data structure and monitors the way of the stored data protection and the data handling method.

MySQL SW consists of the server, a few programs for MySQL DB administration, and supporting programs. Server MySQL manages all tasks required by the user from the DB, for example creation of a new DB. It also allows for data addition, retrieval, etc. Thus, it doesn´t differ from the DB environment of your computer in the terms of function, as described above. The difference is in MySQL „living“ on the Internet, i.e. it is operated through web interface. When is it suitable for use? Of course, originally it was developed for solution of tasks where more people wanted to work with the same data. This is impossible if you have the DB exclusively in your PC. On the other hand, it requires much communication and it was a problem at the times of slow Internet. Gradually, communication speeded up, the data transmission is now practically unlimited and MySQL use has become a standard.

Table No. 1 contains overview of the server properties that rank it definitely among the first-choice options when deciding on DB systems.

Taking in account all advantages presented in the table, one should consider the limitations resulting from the fact that the systems MySQL are not intended for interactive development of applications, like Access or FileMaker. To fully utilize all its functions, we should be familiar with the ways of user environment creation on the web and this requires knowledge and experience in work with PC language HTML and at least the principles of PHP. Regular clerk working in the public healthcare sector would probably not study it all for creation and use of a simple DB application. MySQL use for basic DB work in the terms as shown on the mortality example is fully available without the need for deeper study of the web application programming.

- *Fast*. MySQL authors were mainly aimed at providing for fast speed.
- *Cheap*. MySQL is available free-of-charge under the open source license GPL and the commercial license price is very reasonable.
- *Simple*. You can create and work with DB MySQL with a few simple orders in computer language SQL, a standard language for communication with RDBMS.
- *Flexible*. MySQL runs on various operating systems - Windows, Linux, Mac OS, majority of Unix versions (including Solaris and AIX), FreeBSd, OS /2, IRIX and others.
- *Widely available.* It is available on almost all web-hosts, i.e. widely available at no further cost.
- *Technical support*. Many users use free-of-charge support via e-mail conference.
- *Safe.* MySQL enables particular users/ user groups to create a flexible system of permits and DB rights (for example DB creation privilege or data deleting privilege). Passwords are enciphered.
- *Support of large DBs*. MySQL processes min. 50 million of rows. The table file limit refers to 4GB but it is possible to theoretically increase it to 8 GB if your operating system is capable of managing it.
- *Adaptable*. Open source GPL license allows the programmers for MySQL SW modification.

Table No. 1 MySQL DB properties pursuant to [9].

Summary

In the first part, we presented the basis of data storage technique. We discussed not only technical tools but mainly the principles of PC technique functioning. The second part is dedicated to DB systems and their application to data processing in our area of interest, i.e. public health. Use of a simple DB system was illustrated on the example of mortality data taken from WHO. At the end of the chapter, we presented MySQL characteristics as a system that is both freely available and allows for creation of web applications.

References

1. Magner, L. N. *A History of Medicine*. 2005: Taylor & Francis Group, LLC.
2. Doll, R. et al. *Mortality in relation to smoking: 40 years' observations on male British doctors*. BMJ 1994.
3. Balážová, Ľ., Bosak, J., Genzor, J., Ripka, I., Skladaná, J. ed. *DICTIONARY OF ADOPTED FOREIGN WORDS (ACADEMIC)*. 2005, Slovenské pedagogické nakladateľstvo, Mladé letá, s. r. o.: Bratislava.
4. TechTarget. *SearchSQLServer*. [cited 2013 25.03.]; Available from: http://whatis.techtarget. com/.
5. LibreOffice. *Introducing Base*. in *Base Guide*. 2011, The Document Foundation.
6. The Document Foundation. *LibreOffice in Slovak language.* [cited 2013 25.03.]; Available from: http:// sk.libreoffice.org/libreoffice-v-slovencine/
7. WHO. *WHO Mortality Database*. 2012 [cited 2013 25.03.]; Available from: http://www.who. int/healthinfo/mortality_data/en/index.html
8. *MySQL. The world's most popular open source database.* [cited 2013 17.05.]; http://www.mysql. com
9. Valade, J. *PHP & MySQL® For Dummies*. 4th ed. 2010: Wiley Publishing, Inc., 111 River Street, Hoboken, NJ 07030-5774.

Chapter 5

Data sources and healthcare IS

Chapter objectives

In this chapter, we would like to present the data sources that we drew the examples stated herein from or which are frequent data sources for the public health identification. Let´s explain the term *data source* in detail. What´s behind it? In the book about bio-medicine statistics [1], we enlisted the data sources for research, focusing on those obtained experimentally by measurement. In this book, already processed and then published figures represent the main source. Thus we will not describe the details of data origination, acquisition or collection method, despite of the need for doing it from time to time, but we shall focus on the sources with possibility to obtain data files.

We intend to present available sources of demographical data, data about diseases, and public health sources from domestic and foreign institutions. We will try to present the most important ones in this area, whose statistical data are used in the examples in the following chapters.

If necessary, we stated also the access path to the provided data. We will present the examples of use of DB systems, table editors and mainly of {R} environment functions. Of course, we are not capable of presenting all data sources and we have to select from a wide spectrum thereof. We took various criterions in account. Speaking of domestic sources, we focused on the official statistics accessible at the *Statistical Office of the Slovak Republic* and the publications from the *national Center of Healthcare Information.* As for international sources, we chose from those where Slovakia makes regular contributions to, e.g. international organizations associating world countries, e.g. *WHO – World Health Organization.* We also presented in detail the institutions providing data about regional associations that Slovakia is a member of, e.g. EU sources represented mainly by the organization *Eurostat*. We enlist also some treaties and standards that regulate the data collection, processing and distribution on national or international level, for example ISO standard regulating time data processing.

One of the chapter objectives is to point out the sources, present their general map and clarify the principles of data access. In the following chapters where we work with the data, it will not be necessary to come back to the basic information.

Foreign and domestic sources

Official statistics

Official statistics represent a fundamental source. This expression is used for the data collected and distributed by many national or international agencies. [2] Data are collected in order to provide quantitative grounds for decision-making on the strategies, policies and research support. This is also a topic of this publication. The term agency includes various institutions. In some countries, statistical offices discharge this function; in Slovakia it is the Statistical Office of the Slovak Republic. In other countries, it is the system of governmental organizations; all of them collect and process a part of national statistics. Economic data are processed by the Central Bank. Moreover, particular Ministries contribute to the statistics (Ministry of Healthcare, Education, Agriculture, Social Matters, etc.) with the data collecting for their own needs. National organizations are usually obliged (legal duty) to provide the agreed data to international organizations. They are either on the regional level, like Eurostat, dealing mainly with EU member countries; or on the global level, like United Nations, collecting and providing statistics from the whole world. Along with UN, there are other globally performing international institutions, for example WHO in the area of health, Food and Agriculture Organization in the area of nutrition and food sector, and the World Bank in the area of finances and economy.

Activity performance

All the mentioned organizations share data, prepare reports on the data basis and distribute them. If you visit the WHO website or web pages of WHO branch office in Slovakia or any of its central offices, you can find publication everywhere that explain certain aspect of WHO interests.

The same applies to other international organizations and, last but not least, to the institutions dealing with the statistics on the national level. We can generalize: national and international sources of statistical data are generally available from the Internet or in printed form. At least once per year, they are updated in the form of summary statistical reviews. However, it is more complicated on lower levels, e.g. on the regional, city or district level in case of Slovakia. Certain data can be obtained only upon request from the competent offices.

Statistical standards

Neither national nor international statistics is possible without the treaties on statistical data exchange, concluded on multinational level. International Classification of Diseases (ICD) is well known by healthcare professionals [3] as an example of death cause standardization. We describe it in a separate chapter together with further classifications.

ISO 8601

Here we would like to pay attention to the standard included in so called ISO standards. Standard No. 8601 specifies the method of working with certain statistical elements and their formats. [4] In the introduction, it´s written that the international standard is aimed at eliminating the risk of incorrect interpretation and preventing from uncertainties and their consequences. This standard contains numerical representation of information related to time, e.g. dates, weeks, time intervals, combinations of date and time and time information during the course of a day. These specifications apply exclusively to the Gregorian calendar and 24-hour day. Overview of the items determined by the standard is contained in the table No. 1.

- Calendar date expressed by calendar year, calendar month and calendar day in a month;
- Ordinal date expressed by calendar year and calendar day in a year;
- Week date expressed by calendar year, calendar week No., and calendar day in a week;
- Local time based on 24-hour time monitoring system;
- Coordinated Universal Time in the day;
- local time and the difference from the Coordinated Universal Time;
- Date and time combination within a day;
- Time intervals;
- Repeated time intervals.

Table No. 1 Overview of items regulated by standard ISO 8601 [4]

As a standard example, let´s talk about the calendar date expression. Calendar year is expressed with four digits. Calendar years are numbered upwards according to the Gregorian calendar within values (0000) - (9999). The values from (0000) to (1582) should be used according to mutual agreement of partners on the information exchange. Calendar month is expressed by two digits. January refers to (01) and the following months are numbered upwards. Calendar day in a month is expressed with two digits as well. The first calendar day of a calendar month is expressed as (01) and the following calendar days of the same calendar month are numbered upwards.

If a calendar day should be presented in a complete mode, it should be expressed through one of numerical formulas where (YYYY) refers to calendar year, (MM) refers to calendar month in the calendar year, and (DD) refers to order No. of calendar day in the calendar month.

Basic format:	YYYYMMDD
Example:	19850412
Extended format:	YYYY-MM-DD
Example:	1985-04-12

Further details are contained in the standard. We are not going to talk about it in details; the standard is available in English language. The standard has been more and more used in data communication and is legally binding for all EU countries. It is an example of standardization that applies also to the statistics.

Public Administration IS

The Slovak Ministry of Transportation, Posts and Telecommunications, Section of Company Informatics, Department of Information Security and Standards, issued the *Methodical Guideline to Public Administration IS, its Structure and Requirements, Method of Update and Public Administration IS Application Areas. [5]*

The guideline provides extended information on the structure and requirements of the standards, method of their updating and application areas of the public administration IS; as the tools of information activity implementation during development and operation of the public administration IS or its part. The areas of standardization regulated by the guideline include technical standards, accessibility standards, electronic data exchange, standards of electronic services terminology, and safety standards. The list isn´t definite and will be extended in the form of further guidelines that should include for example data standards, spatial identification standards, etc. Application of the standard called *Statistical Data and Metadata Exchange, version 2.0 (SDMX 2.0)*. is recommended/ required during the statistical data and metadata exchange. [2]

SDMX

Through common efforts and supported by the Bank for International Settlements (BIS), European Central Bank (EcB), Eurostat, International Monetary Fund (IMF), OECD, UN, and the World Bank focused on the development of more effective process of data and metadata provision and exchange within their competences. They jointly created SdMX [2], representing technical and statistical standards and guidelines, together with IT architecture and IT tools to be used for efficient exchange and provision of statistical data and metadata. Standardized data file formats serving for data and metadata processing represent a precondition of automatic production, processing and exchange of SdMX data and file metadata between national and international statistical organizations. With regard to the domains where these standards are mainly used, these include also demographics, education and epidemiology, along with financial and banking sector. The standards are strictly technical and this publication is not aimed at describing them in detail; since they are mainly intended to technicians creating information systems /IS.

Data exchange formats

Increasing data exchange among various institutions resulted in the need for standardized data description so as they can be exchanged among various platforms (e.g. DBs or table editors regardless the origin). One of them refers to the language XML, namely Extensible Markup Language, developed and standardized by consortium W3C (World Wide Web Consortium) as a follow up of the language SGML and generalization of language HTML. [6] It allows for simple creation of particular programming languages serving for various purposes, and a broad spectrum of various data types. The language is mainly intended for data exchange among the application and publishing of documents. It enables description of a document structure in the terms of factual content of the parts but doesn´t deal with the image of the document or its part. The image is then defined through the attached style. Another option is transformation to other document type or other XML structure through various styles. In its package Microsoft Office 2007 and the newer versions, Microsoft stores all files in this format, adding letter *x* in the extended file name, for example *.docx*, *.xlsx* a *.pptx*. These formats and suffixes of the files are used in the programs Microsoft Word, Microsoft Excel and Microsoft PowerPoint.

Text file marked as *CSV*, i.e. *Comma-separated values,* is frequently used format. [7] It is a file containing plain text without formatting marks used by text editors or other programs. It is intended for storage of table data. File in this format consists of any amount of entries/ rows separated by a new row mark. Each entry/ row contains columns separated by other mark, usually by comma (,) or tabulator.

All entries have usually similar number of columns. This format is commonly used at exchange of large data amount. In this way, information can be transferred for example between table editor and DB program or {R} environment. The first row usually contains the names of columns that the receiving program, for example {R} interprets as the names of variables.

Many organizations offer the data directly in the format intended to table editors, or any of the DB formats.

Dictionaries

At the end of this section, we would like to make you aware of the existence of more standards than stated herein. We can add terminology to the standards as well, for example the epidemiology dictionaries [8, 9]. The first one was translated in the Slovak language as well, [10] similar to the demographical dictionary. [11] EUROSTAT provides for a very good service through website *Statistics Explained*. [12] It is a portal with additional information for occasional and regular users, a cyclopedia of the EU Statistics and last but not least, also the dictionary of statistics. It contains statistical data and explanations as well as direct reference to the latest data and all types of related information that can be useful for understanding the statistics. The portal core refers to statistical articles. Each Eurostat statistical file is described with one or more articles. Typical article starts with specification of timeliness of the data used, continues with a brief introduction and detail article text.

Domestic routine statistics

Legal standards

The *State Law of Statistics* [13] represents the legal basis regulating the terms of data and information acquisition, required for evaluation of social- economic development, position and competence of the authorities providing for the state statistics, tasks of the public authorities in the area of the state statistics, rights and obligations of intelligence units, protection of confidential statistical data against misuse, provision and publishing of statistical data, provision for comparability of statistical information and fulfillment of commitments resulting from international treaties in the area of the state statistics that the Slovak Republic is bond by. The law describes the main terms as well.

ŠÚSR

The Slovak Statistical Office fulfills tasks of the state statistics. Ministries and the state organizations are allowed to fulfill the state statistics tasks upon the terms and in extent specified in the above law. We should mention public administration IS as a source of statistical data, providing for acquisition, provision and accessibility of data, collecting, processing, transfer, storage, archiving and disposal of data within the law principles. [14]

Statistics of Health

Statistics covers two areas of health – healthcare statistics and work incapability statistics in the area of disease and injury. Data is acquired in the area of healthcare statistics within the regular state annual monitoring performed by the Slovak Ministry of Healthcare. [15] They are data related to outpatient´s department and hospital care, joint therapeutic and examination files, information on the chain of healthcare facilities in Slovakia and on health condition. NCZI possesses also further information on the population health condition, obtained from statistical monitoring based on the legislation of the Slovak Ministry of Healthcare. Further health-related data sources represent the administrative sources of the Slovak Public Healthcare Office, the State Institute of Pharmaceutical Control, Health Insurance Companies and the Social Insurance Company. Methodology descriptions also provide much useful information, containing fundamental characteristics of each of the indicators for the area of healthcare system and work incapability.

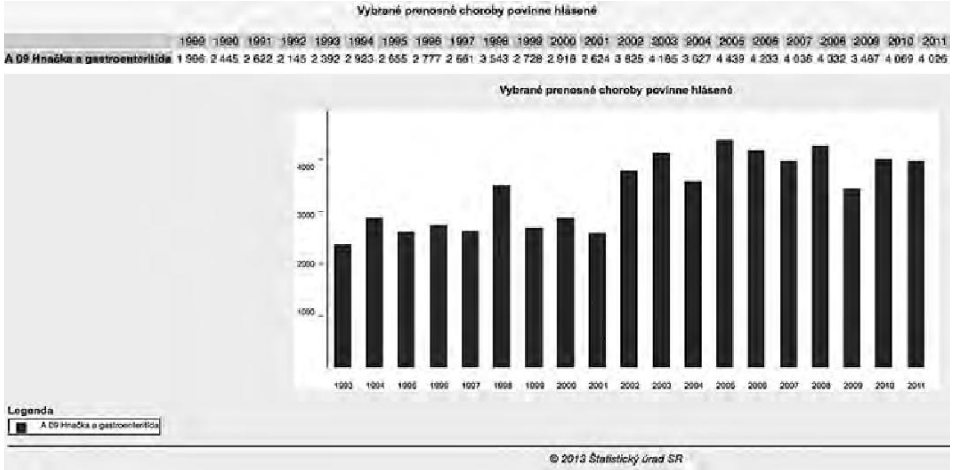
SLOVSTAT

The Slovak Statistical Office offers the data via DBs available on the Internet. Data can be requested for purpose of special statistical analysis, which are not commonly accessible. DB SLOVSTAT [16] is a database intended for the amateur public and the professionals, containing the latest information and time lines of the indicators of social – economic development of the Slovak Republic. The DB makes accessible fundamental data from all statistical sections in monthly, quarterly or annual lines. Many indicators are structured in detail. The data are updated on regular basis and made available in a short time interval after official publishing. In case of revisions, all data time lines are modified. The DB contains extensive amount of the tables from particular statistical sections, thus they can be accessed through the navigation tree that is dynamically opened and closed upon click on the thematic and statistical circle hyperlink down to the lowest navigation level with the list of selected statistical tables. Statistical topics create the basic navigation tree axis, for example demography and social statistics, business statistics or multi-purpose statistics. Each topic is further branched to statistical areas in which the statistical tree branches contain detail statistics of particular area and corresponding time lines. The table name determines the area of displayed indicators and the time line in detail. They are created dynamically according to indicator and time selection. The data are displayed therein in rows as a time development of the indicator value, where the columns specify the time of particular value. “Text menu” can be also selected (on the top toolbar), offering unpacked navigation tree with overview of provided statistics. The application contains abbreviations and used signs, operating manual and fast search based on various criterions, as a supporting tool for the users.

- Statistical infrastructure
- Demography and social statistics
  - 0 Demography and social statistics
    - ■ Demographical statistics, methodology
    - ■ Characteristics of citizens´ movement
    - ■ Demographical characteristics
    - ■ Citizens´ age structure
    - ■ Citizens´ structure according to marital status
    - ■ Foreign migration
  - 0 Labor market
  - 0 Household income, expenses and consumption
  - 0 Educational system
  - 0 Health
    - ■ Selected healthcare indicators
      - – Jobs and beds in healthcare facilities (2005 – 2010)
      - – Physicians in healthcare facilities and bed fund in institutional healthcare according to hospital wards as of Dec 31. (1995 – 2010)
      - – Pharmacies, consumption of pharmaceuticals (1995 – 2011)
      - – Selected contagious diseases obligatorily reported (1989 – 2011)
    - ■ Work incapability
    - ■ Financial indicators in healthcare sector
    - ■ Historical data
  - 0 Social security
    - ■ Retirement Plan and sickness insurance
    - ■ State allowances
    - ■ Social aid
    - ■ Social protection
    - ■ Historical data
  - 0 Culture
  - 0 Criminality, fires
- Macroeconomic statistics
- Business statistics
- Agriculture, forestry, fishing
- Monetary and financial sector, foreign trade
- Multiple statistics

Table No. 2 SLOVSTAT DB structure [16]

Health statistics doesn´t naturally create the main part of offered statistics. It mainly contains aggregated data that can be used for more detail study in limited extent. Healthcare data are published in aggregate for the Slovak Ministry of Healthcare, Slovak Ministry of Defense, Slovak Ministry of Internal Affairs, Slovak Ministry of Justice, and Slovak Ministry of Transportation, Posts and Telecommunications. The indicators contain brief characteristics. Example of provided data is on the picture No. 1. Selected data can be exported to Excel or in the form of text separated with comma for further processing, or printed. Selection of variables is simple and intuitive.



Picture No. 1 Data related to diarrhea and gastroenteritis incidence in Slovakia within 1989 – 2011in the form of table and graph. (Source: SLOVSTAT)

NUTS

The Slovak Statistical Office provides the data about the Slovak regions, cities and towns. Regions are delimited on basis of classification called „Nomenclature Units of Territorial Statistics" (French abbreviation NUTS). It was introduced by the Statistical Office of EC (Eurostat) in cooperation with national statistical institutes; in Slovakia with the Slovak Statistical Office. NUTS structure doesn´t have to necessarily correspond to the state administration structure. NUTS levels are specified in Slovakia (as statistical units) on basis of the Slovak Statistical Office measure and its acceptation by the Eurostat as follows: NUTS I – Slovak Republic; NUTS II – 4 statistical territorial units: Bratislava region, Western Slovakia (region Trnava, Trenčín and Nitra), Central Slovakia (region Žilina and Banská Bystrica), Eastern Slovakia (region Prešov and Košice)/; NUTS III – eight self-governing regions (VÚC); NUTS IV – 79 districts; NUTS V – 2883 municipalities. Levels NUTS 4 and NUTS 5 used in the past were renamed to LAU 1 and LAU 2 (Local Area Units) (Table No.3).

- Slovak Republic (NUTS 1)
- Territory (NUTS 2)
- Region (NUTS 3)
- District (LAU 1) corresponds to (NUTS 4)
- Community (LAU 2) corresponds to (NUTS 5)

Table No. 3 Overview of statistical territorial units in Slovakia

RegDat

Regional Statistics DB (RegDat) contains time lines of the economic and social development indicators in particular Slovak regions (regions, districts and territories, based on the data availability). Regional data from various statistical areas are in the form of tables in various annual time lines. Their structure is very similar to the previous one. Data access is enabled through DB tool PX-WEB. To illustrate available outputs, we chose data selection from city Bratislava divided in five districts. We chose work incapability statistics in the area of illness and occupational injury for both genders in year 2001 - 2011(Picture No. 2.

	Average work incapability percentage		
	2010	**2011**	**2012**
TOTAL			
District Bratislava I illness occupational injury	 1,752 0,020	 1,647 0,010	 1,510 0,010
District Bratislava II illness occupational injury	 2,112 0,020	 1,964 0,017	 1,933 0,021
District Bratislava III illness occupational injury	 1,965 0,011	 1,821 0,023	 1,802 0,022
District Bratislava IV illness occupational injury	 1,714 0,023	 1,740 0,024	 1,575 0,023
District Bratislava V illness occupational injury	 2,151 0,022	 1,142 0,025	 2,132 0,020

Methodical sheet. Work incapability data have been taken over from the Social Insurance Company since 2007. Indicators for Slovakia in 2007 include the cases of insured persons without territorial affiliation. Data from 2007 neither include the structure according to genders.

Latest update: Aug 19, 2013

Picture No. 2 Work incapability in Bratislava districts (Source: SOURCE: STATISTICAL OFFICE , 2013)

DB allows for data presentation in various formats. Simple data selection in combination with possible time line export allows for data application to detail research in the area of public health.

MOS and MIS

Information system (IS) of the City and Community Statistics (MOS) [17] provides for SU SR requirements on central, regional and local branch level in the area of data collection, and uploading of indicators for particular Slovak municipalities and cities. MOS IS provides for statistical information per each city and community including possibility to define the Slovak territorial units in the outputs and with perspective interconnection to the EU statistics. It is a system with intersectional nature that is not covered by the state statistics with its findings. Indicators from the city IS (MIS) are defined there in parallel, used in bigger Slovak cities. The system has been implemented so far only in the Slovak capital city Bratislava and in Košice since 2003.

MOS DB [17] that is updated on annual basis, containing data for year 1993 and continuously for all years afterwards since 1996. Data are collected annually by SU SR branch offices. The DB is kept as a central DB containing data from all Slovak municipalities and regional DBs operated at SU SR regional offices for the regions and municipalities in their competence. The list of the areas and indicators in MOS DB is stated in the table No. 4.

All data are presented either as formatted in the table or a plain text. Indicator 7 – healthcare and curing – provides statistical data about policlinics, physicians´ offices,

01 – basic characteristics,	15 – social area,
02 – transportation,	16 – organizational structure of economy,
03 – technical amenity,	– subjects according to selected economic activities,
04 – demographics,	
05 – habitation fund,	18 – competence of state administration, judicial and insurance bodies,
06 – sport,	
7a – healthcare,	19 – water flows,
7b – curing,	20 – air pollution,
08 – trade and restaurants,	21 – climatic conditions,
09 – lodging facilities,	22 – geological conditions,
10 – banking and insurance sector,	23 – other environmental components,
11 – educational sector,	24 – selected results of *Population Census* in 1991,
12 – culture,	
13 – cultural monuments,	25 – selected results of *Population Census in* 2001.
14 – community acreage,	

Table No. 4 System of indicators monitored within MOS healthcare facilities, pharmacies, first aid, vehicles and spa facilities. Indicator No. 4, demography, includes basic demographical data about newborns, deaths, citizens´ headcount, migration, marriages and divorces. Definition of each item is displaced upon keeping the cursor on particular indicator name or value in the form of indicator description. It supports uniform interpretation and identification of indicators monitored by MOS. The system doesn´t offer the graphical illustration. As an example, we will display healthcare services provided in the town Trnava (Picture No. 3).

The data can be used for regional or local analyses and research.

Trnava Návrat na okres Trnava

Demografia (31.12.2012)

Ukazovateľ	Hodnota
Počet obyvateľov k 31.12. spolu	66073
muži	31940
ženy	34133
Predproduktívny vek (0-14) spolu	8866
Produktívny vek (15-54) ženy	19193
Produktívny vek (15-59) muži	21827
Poproduktívny vek (55+Ž, 60+M) spolu	16187
Počet sobášov	328
Počet rozvodov	164
Počet živonarodených spolu	687
muži	341
ženy	346
Počet zomretých spolu	524
muži	291
ženy	233
Celkový prírastok (úbytok) obyv. spolu	-146
muži	-116
ženy	-30

Zdravotníctvo (31.12.2012)

Ukazovateľ	Hodnota
Lekárne a výdajne liekov	nie
Samostatné ambulancie praktického lekára pre dospelých	nie
Samostatné ambulancie praktického lekára pre deti a dorast	nie
Samostatné ambulancie praktického lekára stomatológa	nie
Samostatné ambulancie praktického lekára gynekológa	nie

Životné prostredie (31.12.2012)

Ukazovateľ	Hodnota
Skládka komunálneho odpadu	áno
Komunálny odpad	áno
Využívaný komunálny odpad	áno
Zneškodňovaný komunálny odpad	áno

Picture No. 3 Demographical, healthcare and environmental indicators for the town Trnava
(Source: MOŠ, 2013)

VDC

Research demographical center (VDC) [18] was founded in 2000. It performs within the Institute of Informatics and Statistics in Bratislava (InFOstat). VDC activity is focused mainly on the assessment of population development and its parts; preparation of population estimations and prognoses and preparation of analytic and prognostic grounds for decision making by top managing authorities in the area of social, economic and habitation policy, employment, healthcare and education. It also fulfills tasks in the area of methodic and methodology, dealing with preparation of methods, models and tools for demographical analyses, estimations and prognoses, update and operation of demographical IS and development of demographical methodic and methodology, among other tasks. It also coordinates demographical research in Slovakia. In the terms of health study and health-related services, DBs on the Slovak population are important as well as death cause DBs. VDC is mainly oriented to demography, thus the data offered are exclusively of demographical nature. Data about population, its movement and death causes are important for those interested in information on the population health.

- Tables of basic demographical data in Slovakia
- Movement of Slovak citizens (detail demographical data)
- SLOVAKPOPIN website
- Slovak population prognosis tables 2003 (2005-2050)
- Slovak population prognosis tables – areas - 2003 (2005-2025)
- Slovak population prognosis tables – regions - 2003 (2005-2025)
- Slovak population prognosis tables- districts - 2003 (2005-2025)
- Slovak population prognosis tables- regions - 2008 (2008-2025)
- Slovak population prognosis tables – districts - 2008 (2008-2025)
- Slovak population prognosis tables 2012 (2012-2060) – low, medium and high variant
- Slovak death tables
- Life median at birth – Slovak regions, Bratislava city, Košice city
- Life median at birth – Slovak districts
- Life median in health based on EHIS2009
- Death cause categories according to Slovak areas (NUTS2) and Slovak regions (NUTS3)

Picture No. 4 Items from DB available via VDC
(Source: website VDC [18])

The website offers the tables but transformation in the form suitable to further processing. It contains the tables of prognoses and life medians, and offers link to website SLOVAK POPIN [19].

SLOVAK POPIN

Population Information Network, POPIN was founded by UN in 1979, namely its Economic and Social Board. The organization is aimed at making accessible international, regional and national information on the population. Slovak Republic is a member of the group, actively contributing to fulfillment of the goals through data provision. SLOVAK POPIN is in English language as well as all documents downloadable from the website. They are divided in typical demographical categories (Picture No. 5).

Population statistics	Divorce rate
General information	Abortions
Major demographical indicators	Birth rate
Age structure	Death rate
Marriage rate	Death causes
	Migration

Picture No. 5 SLOVAK POPIN DB contents (Source [19])

They are all suitable for the study of the public health, nevertheless the death statistics is probably most frequently discussed topic. The structure of freely accessible data is available in three forms: files *XML*, *XLS* (MS Office Excel 2003) or *HTML*. The data in the first format are primarily intended for storage in DB environment. XLs are data primarily readable through the program Excel. The last format is intended for table browsing on the display and eventual transfer via function Copy-Paste.

Reading *CSV* files is easiest for {R} environment. Authors of {R} environment don´t recommend reading *XLS* files because of too complicated specifications of the function arguments, however it is possible. It is better to export the data to *CSV* format and then to {R}. [20] Data transfer via clipboard can be use as well. At first, we will demonstrate *CSV* file creation from the file *XLS* (Example 1) in the selection from DB POPIN. We will obtain the data from *Main Demographic Indicators* in *XLS* format that should automatically open Excel file. Thereof, we will choose those of our interest and transfer them to empty Excel page. Then we save them as a text file with suffix *.csv*. This can be read in {R} environment. In the R Commander environment, it is necessary to enter the name of variable in which the data should be stored. It will be created in the form of data array. We should also confirm that the names of variables are in the first row. Since the file is read from the PC disc, we should state it. We can see used semicolon as a data separator. We work with data that are whole digits but in case of decimal digits, we should state that decimal point shall be comma or dot.

	A	B	C	D	E	F
2						
3	Tab.	Main demographic indicators				
4		1950	1951	1952	1953	1954
5	Population (as of 1st January)	3447085	3485530	3533282	3576852	3629425
6	- males	1671836	1688740	1712071	1732792	1761164
7	- females	1775249	1796790	1821211	1844060	1868261
8	Population (as of 1st July)	3463446	3508698	3558137	3598761	3661437
9	- males	1678970	1700135	1724275	1744174	1779490
10	- females	1784476	1808563	1833862	1854587	1881947
11	Main age groups (as of 1st January) - males					
12	0-14	499540	509544	521450	536009	552504
13	15-59	1025525	1031702	1041093	1046007	1055347
14	60+	146771	147494	149528	150776	153313
15	0-19	653259	659781	669282	681206	697911
16	20-64	920090	929954	942083	949918	960151
17	65+	98487	99005	100706	101668	103102

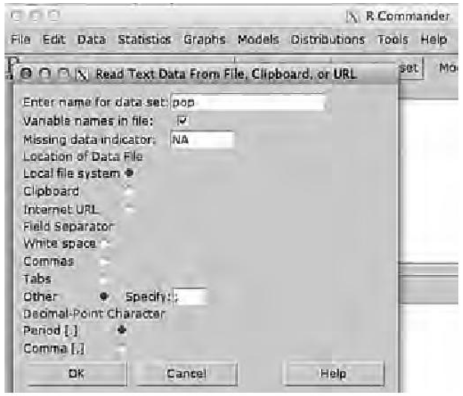
Data chosen and entered as values (otherwise zero is transferred)

1994	1995	1996	1997	1998	1999	2000
5336455	5356207	5367790	5378932	5387650	5393382	5398657

Entry data after uploading to Excel

```
1994;1995;1996;1997;1998;1999;2000;2001;2002;2003;2004;2005;2006;2007;2008;2009;2010;2011
5336455;5356207;5367790;5378932;5387650;5393382;5398657;5378783;5378951;5379161;5380053;53
84822;5389180;5393637;5400998;5412254;5424925;5392446
```

File in the format .CSV, the 1st row contains names of columns and the 2nd row contains the values




```
> pop <- read.table("/Users/martinrusnak/Desktop/pop.csv", header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE
> pop
    X1994   X1995   X1996   X1997   X1998   X1999   X2000   X2001   X2002   X2003   X2004   X2005   X2006   X2007   X2008
1 5336455 5356207 5367790 5378932 5387650 5393382 5398657 5378783 5378951 5379161 5380053 5384822 5389180 5393637 5400998
    X2009   X2010   X2011
1 5412254 5424925 5392446
```

Result – data read in R environment

Example 1 Data reading from POPIN to {R}environment

We shall send the request and specify the file that we intend to read from. If we have made correct request, we should obtain the variable containing the digits from DB.

Then, in the similar way, we upload data from the table displayed by the browser after chosen format HTML. The required data should be removed from the box through the order Copy or Ctrl-C and continue like in the previous example.

T 2.8.2 PACIENTI V PRAVIDELNEJ DIALYZAČNEJ LIEČBE (PDL) PODĽA ZÁKLADNEJ DIAGNÓZY A ÚZEMIA ZDRAVOTNÍCKYCH ZARIADENÍ

PATIENTS IN REGULAR DIALYSIS TREATMENT (RDT) BY DIAGNOSIS AND TERRITORY OF HEALTH ESTABLISHMENT

Diagnóza *Diagnosis*	Počet pacientov v PDL / *Number of patients in RDT*								
	SR	BL	TA	TC	NI	ZI	BC	PV	KI
Primárne glomerulonefritídy *Primary glomerulonephritis*	**576**	94	57	51	89	76	57	69	83
Pyelonefritída *Pyelonephritis*	**615**	67	77	64	66	81	65	117	78
Polycystická choroba obličiek u dospelých (dominantný typ) / *Polycystic kidney disease by adults (dominant typ)*	**224**	21	19	24	36	36	35	28	25
Poškodenie obličiek hypertenziou *Injury of kidney by hypertension*	**327**	56	30	46	27	37	63	10	58
Renálne vaskulárne ochorenia *Renal vascular diseases*	**207**	24	16	18	38	34	27	9	41
Poškodenie obličiek spôsobené diabetes mellitus *Injury of kidney due to diabetes mellitus*	**1 351**	121	171	170	154	148	175	207	205
Neznáma *Unknown*	**115**	26	12	25	10	14	1	25	2
Iná *Other*	**637**	49	44	78	75	73	91	117	110

We transferred them to Word with Copy-Paste function and deleted the text before the rows; we also transferred the column headlines.

We would like to continue working with the table data

> kidney <- read.table(“clipboard”, header=TRUE, sep=””, na.strings=”NA”, dec=”,”, strip.white=TRUE)

We uploaded the clipboard removed from Word to R environment

> kidney

```
  SR BL TA TC NI ZI BC PV KI
1 583 78 78 48 64 70 68 95 82 2
704 70 82 84 57 92 107 127 85 3
237 21 19 30 38 41 39 26 23 4
294 48 27 30 23 30 57 15 64 5
206 21 14 30 30 34 29 11 37
6 1297 100 169 187 142 134 167 199 199
 7  86 23   8 15   7 15   3   5 10
 8 594 79 34 43 101 73 65 100 99
```

Example 2 Reading data from the Statistical Almanac to the variables in {R}environment

The only difference refers to marked *Clipboard* in the window *R Commander*. Saving on the disc and selection of the file is not necessary, other steps are similar.

NCZI

National center of healthcare information (NCZI) is the state contributory organization dealing with collection, processing and provision of health statistical data. [15] It is an organization, a main provider of health data and healthcare services in the Slovak Republic. The Almanac Journal with annually published *Slovak Health Almanac* represents the basic product hopefully known by everybody dealing with the public health. [21] Website contains volumes starting from 1996 in the format Adobe *pdf*. This imposes also the data usage restriction. Transformation from this format to the one suitable for further processing is rather time demanding. Many data are only available in the form of graphs. Rewriting in the table editor sheet represents one of the options of data usage. It is not very elegant solution. Another solution would refer to utilization of the R function ability *read.table()* to read the clipboard content (Example 2). Take care! They are separated in the tables with spacing but the spacing should be deleted in order to be able to read it properly, otherwise {R} can recognize it as two values and generate error message because of incorrect number of columns. Uploading result will enter the data array in the variable class. Handling with it is easier through the function *attach()*, allowing for work with particular variables.

Creative reader will definitely find also other ways of data transfer from the almanacs to the statistical programs environment, table editors or DBs.

Further interesting data (also in format *pdf*) are available in the section *Edition of Analytic Publications [21]* Here you can find the data about oncology diseases incidence or development of injuries, and further publications.

National Health Registers also represent an interesting source of data. NCZI defines them as *healthcare IS primarily aimed at collecting and processing the data about serious mass diseases, and providing the professionals with basic epidemiological and clinical characteristics related to the diseases occurrence and causes.*[22] The following table No. 5 provides an overview of registers under control of NCZI. Some of them provide the data only upon substantiated request. The other ones, for example The Oncology Register, make publications freely available with the data tables in the format *pdf*.

national register of basic health data
national register of oncology patients
national register of patients with 1st type diabetes mellitus
national register of patients with congenital heart disorder
national register of patients with cardiovascular disease
national register of patients with cerebrovascular disease
national register of patients with chronic lung disease
national register of patients with congenital development defects

Table No 5 national health registers under NCZI control

Out of NCZI control, there is the national Transplantation Register, national Register of Patients with Tuberculosis, national Register of Patients with Transmitted Diseases and national Arthroplastic Register.

International sources of statistical data

Availability of valid, reliable and comparable health information that allow for creation of local, regional, national and global strategies can be fulfilled through four mutually interconnected initiatives: improvement of technology and population health condition measurement methods through strengthened national capacities and governmental motivation to collect and analyze useful data about health condition, setting forth global standards applicable to basic health indicators and their measurements; and provision of globally-valid, reliable and comparable reports enabling evaluation of inputs and results of health service provision. [23] International sources of data in fact copy the system of national statistics with the only difference – they provide for overview of data from various countries. They are aimed at providing comparable material to the countries and are usually less detail than the national statistics.

MacroDataGuide

MacroDataGuide [24] is a user manual for researchers and students, globally used when searching for quality data for purpose of social-scientific research. It provides information on the data availability and quality assessment of these sources. It was created and has been kept by the *Norwegian Social Science Data Services* (Nsd), a limited liability company owned by the Norwegian Ministry of Education and Science. The portal refers to the report prepared for the *European Social Survey (ESS)*, investigating the availability and comparability of the existing contextual statistical data sources. These can be interesting for ESS data user. Web portal represents a website presenting information from various sources in a uniform way. Usually, every information source is published on the specific website called *portlet* and a user can choose which one he wishes to display. [25]

MacroDataGuide sources contain evaluation of actual availability and coverage of context data sources. While it was originally intended for the European micro-data users (ESS users), they have gradually extended its reach and today provide macro-data to many countries. The data sources differ in the terms of the extent and specialization and include general DBs containing social and economic data, non-governmental and inter-governmental organizations. Data files are compiled by independent researchers.

Information provided on this portal comes from various institutions and research projects. Moreover, there are independent DB reviews included, published in magazines, and relevant studies. The portal provides for efficient tool that is time- and cost saving for the one who is interested in data sources from the whole world.

Furthermore, we will check three major sources from international organizations. The first one refers to two DBs of WHO, namely *Health for All* and *Mortality database*. The second one represents Eurostat health-related DB.

WHO

WHO prepares and publishes the statistics of a broad range of topics associated with health, offered to the interested and researchers in the form of various DBs and printed publications. Primarily, information on mortality, illness rate, health condition, service coverage and risk factors is provided

European Health for All - EU DB

EU DB called European Health for All (HFA) [26] provides for access to a broad spectrum of basic health statistics to 52 WHO European member countries. The data come from the official registers and surveys. They are collected from various sources. A part of them is collected annually directly in the member countries. Another part of the data comes from WHO technical units that acquire respective statistical information within their own specialization. Secondary sources – data provided by international organizations and agencies represent similarly important source of data. The DB contains approx. 600 indicators arranged in the following groups: demography and social-economic statistics, mortality indicators, illness rate, health disablement and hospitalizations, lifestyle and environmental risk factors, healthcare sources, healthcare services utilization and cost, health of mothers and children. On-line DB version is based on pop-up windows; offline version can be also downloaded. The DB allows for choosing countries by the users, as well as variables and years, when creating the tables. Selection result is displayed either in the form of tables, maps or graphs. All outputs can be printed or exported in offline version for purpose of processing by other tools. The data were compiled, verified and processed in a uniform way in order to improve international comparability of the statistics.

Mortality Database

WHO Mortality DB [27] contains more than 2 million death cause entries in a data file. They include deaths registered in the national death registers in the form of coding by the competent inland authority. The data represent official national statistics submitted to WHO by competent authorities of the countries involved.

Working with data, it is necessary to consider national differences in death registration and coding of procedures, mainly in the usage of codes for incorrectly defined and unknown death causes. WHO provides the data upon requirement not to use them for commercial purposes. Overview of the files incl. contents description is included in the table No. 6.

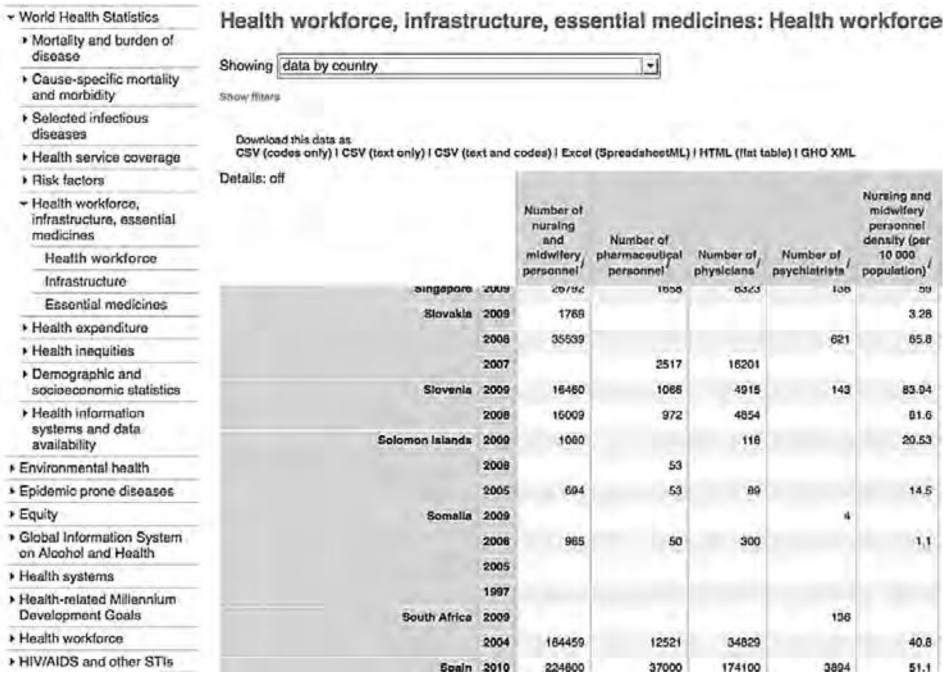
The data are classified in various groups according to years. The file including documentation contains all codes and related explanation, it wouldn´t be possible to work with data without them. The files are formatted as text files with the suffix *.csv* and not intended for untrained user. Data files are primarily intended mainly to research purposes, together with required instructions, file structures and reference tables. Persons using them need DB tools to work with the files since they are too large and complicated for common office programs.

File name	**Content**
Documentation	File formatted in MS Word, containing information on WHO Mortality Database, specifications of files and the list of death causes
Availability	File formatted in MS Excel, containing the list of countries and years to which the mortality and population data are available
Country codes	Country names and codes
Notes	Remarks to some data from particular years and countries
Populations and live birth	Reference populations and statistics of live birth
Mortality, ICD-7	Data file – detail data of mortality for the 7th revision of ICD (International Classification of Diseases).
Mortality, ICD-8	Data file – detail data of mortality for the 8th revision of ICD (International Classification of Diseases).
Mortality, ICD-9	Data file – detail data of mortality for the 9th revision of ICD (International Classification of Diseases).
Mortality, ICD-10	Data file – detail data of mortality for the 10th revision of ICD (International Classification of Diseases).

Table No. 6 Description of contents of files included in WHO Mortality Database

Register of indicators and measurements

Register of indicators and measurements (IMR) is a central source of indicators´ definitions in the form of texts and formats readable by computer. [28] It provides for complete and well-structured metadata about indicators, harmonization of indicators and nomenclature definition management, access to indications definitions via Internet. This ensures harmony between various areas of statistical data. It is supported by the format SDMX-HD. Global Health Observatory Data Repository storage is located within IMR, containing extensive list of indicators that can be chosen according to the topic or through a multiple question. It represents main storage of data, indicators and information of WHO health statistics. The storage basis represents 50 data files about priority health topics, including mortality and illness seriousness, development goals of the millennium (children nutrition, health of children and mothers, reproduction health, vaccination, HIV/AIDS, tuberculosis, malaria, neglected diseases, water and sewage), not transmitted diseases and risk factors, epidemic diseases, healthcare systems, environmental health, violence and injuries, equality, etc. Moreover, online access is provided to the annual health-related data summary to 194 WHO member countries, called World Health Statistics.



		Number of nursing and midwifery personnel	Number of pharmaceutical personnel	Number of physicians	Number of psychiatrists	Nursing and midwifery personnel density (per 10 000 population)
Singapore	2009	26792	1658	8323	136	59
Slovakia	2009	1768				3.28
	2008	35539			621	65.8
	2007		2517	16201		
Slovenia	2009	16460	1066	4915	143	83.04
	2008	16009	972	4854		81.6
Solomon Islands	2009	1080		116		20.53
	2008		53			
	2005	694	53	89		14.5
Somalia	2009				4	
	2006	965	50	300		1.1
	2005					
	1997					
South Africa	2009				136	
	2004	184459	12521	34829		40.8
Spain	2010	224800	37000	174100	3894	51.1

Picture No. 6 Example of selection from Global Health Observatory Data Repository

Picture 6 shows that data can be downloaded in various formats except the table, for example text separated with comma, i.e. *.csv*, or in the format readable by Excel or HTML. This provides actually very wide opportunities of data processing for purposes of the study of various health parameters on the global worldwide level.

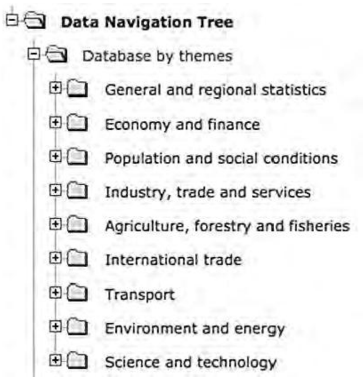
Eurostat

Eurostat [29] is an EU Statistical Office located in Luxemburg. Its role is to provide statistics to EU that allows for comparison of the countries and regions on EU level. Eurostat was founded in 1953 as a support of Coal and Steel Society. With years passing, its function had been extended and Eurostat became one of EC headquarters / directorates general (DG) after EC foundation in 1958. Its key role is to provide statistics for other DGs and data to EU institutions so as they are able to define, implement and analyze EC policies. Eurostat offers many important and interesting data that can be used by government, businesses, educational sector, journalists and the public during their work and in daily life. Accordingly, Eurostat is not a classical statistical institution on the national level. It´s main task isn´t to collect data, this is done by EU member states themselves. It´s main task is to process and publish comparable statistical information on EU level. Therefore its efforts are to establish a common statistical "language" that would include the terms, methods, structures and technical standards. In other words, Eurostat ensures comparability with harmonized methodology.

EUROSTAT DB

Eurostat provides data through its website with well-structured user environment.

It is seen on the picture No. 7 that the access to data is transparent and purposeful. Majority of data can be obtained in various formats– *XLS, CSV, HTML, PDF, PC-AXIS, SPSS, and TSV*. We have been acquainted with the first four. *PC-Axis* [30] is a family of program products consisting of a few programs for Windows and Internet, aimed at presenting statistical information. It is an outcome of worldwide cooperation project of statistics spread. *SPSS* [31] is a commercial package of statistical programs widely



Picture No. 7 Navigation in DB Eurostat used in the science and research. *TSV* format refers to text file similar to *CSV* where particular items are separated by tabulators instead of comma. Along with the data, Eurostat offers metadata describing particular items. It is a useful source of statistical data information.

CODED is a DB of the terms and definitions, and further online dictionaries related to the statistics obtained from surveys. Furthermore, it contains EU legal regulations and methodical manuals related to the statistics; international statistical classifications and nomenclatures, as well as dictionary of the *International Statistical Institute (ISI)* [32] and other on-line synonymic dictionaries related to the statistics. Last but not least, it contains overview of survey methodic used on national levels during preparation of statistics for EU, quality reports, etc. It contains already mentioned Euro- SdMX Metadata Structure (ESMS), in the form of international sets of standards for inter-organizational statistical data exchange. Finally, it contains standard nomenclatures recommended for creation of DBs and data transfer.

Geographical information system (GIS) integrates HW, SW and data for creation, administration, analysis and displaying of all forms of geographically presented information. Eurostat service called GISCO supports and stimulates GIS use within EU Statistical System and EC. It is used for the control and spread of EC geographical reference DB. It produces maps, spatial analyses, supports geo-reference of the statistics and provides user support to GIS users.

Summary

We tried to summarize major data sources of the domestic and international routine statistics. It is impossible to present all sources and it wouldn´t be purposely. Finally, people interested in particular problem to be solved can easily find a suitable source of data. Information in the chapter serves mainly for orientation purposes and we will refer to some of the stated sources in the following chapters. We recommend the chapter reader opening website when reading a source description and trying to browse through the items offered by the product. In this way, one can get more complex view and image of the presented sources and their eventual use.

References

1. Rusnák, M., Rusnáková, V., Majdan, M. Bio-statistics for Students of Public Healthcare System, I. ed, 2010, Trnava: Typi Universitatis Tyrnaviensis, Publishing agency of University TRnava. 216 p.

2. Nelson, C., Gager, J., Gregory, A. SdMX 2.1 User Guide. 2012, Statistical Data and Metadata Exchange SMDX.
3. WHO. International Statistical Classification of Diseases and Related Health Problems, 10th Re-vision, Version for 2007. 2007 [cited 2010 30.12.]; Available from: http://apps.who.int/classifi-cations/apps/icd/icd10online/.
4. ISO, INTERNATIONAL STANDARD ISO 8601. Data elements and interchange formats — Information interchange —Representation of dates and times. 2004, ISO copyright office: Case postale 56, CH-1211 Geneva 20.
5. SR Ministry of Transportation, Posts and Telecommunications, Methodical Instruction on Standards for Public Administration IS and their Structure and Requirements, Methods of Update and Application Areas of Public Administration IS, SR Ministry of Transportation, Posts and Telecommunications, Editor. 2006: Bratislava.
6. Harold, E. R. XML™ Bible. 2004, Foster City, CA, Chicago, IL, Indianapolis, IN, New York, NY: IdG Books Worldwide, Inc. An International Data Group Company.
7. Nassau, C. CSV File Format Specification. 2012, Curzon Nassau: London.
8. Last, J. M. A Dictionary of Epidemiology. 1995, Oxford: Oxford University Press. 196.
9. Porta, M., Greenland, S., Last, J. M. Editors. A Dictionary of Epidemiology. 2008, Oxford University Press, New York: Oxford. p. 289.
10. Last, J. M. Dictionary of Epidemiology. 1999, Bratislava: USAID.
11. Jurčová, D. Vocabulary of demographic Terms. 2005, INFOstat – Institute of Informatics and Statistics. Research Demographic Center: Bratislava.
12. EUROSTAT. Statistics Explained. 2013 [cited 2013 2013-01-17]; Available from: http://epp. Eurostat.ec.europa.eu/statistics_explained/index.php/Statistics_Explained.
13. The State Statistics Act, in: Act No. 298/2010 Coll., 2010: Bratislava.
14. Public Administration IS Act as amended, in: 275, N.R.S. republic, Editor. 2006: Bratislava.
15. National Healthcare Information Center (NCZI). [cited 2013 27.1.]; Available from: http://www.nczisk.sk/O-nas/Pages/default.aspx.
16. slovstat. 2012 [cited 2012 10.1.2012]; Available from: http://www.statistics.sk/pls/elisw/vbd.
17. Statistical IS of MOŠ. [cited 2013 27.01]; Available from: http://app.statistics.sk/mosmis/sk/run.html.
18. Research Demographic Center (VDC). [cited 2013 27.01.]; Available from: http://www.infostat.sk/vdc/sk/index.php?option=com_content&task=view&id=5&Itemid=4.
19. Slovak POPIN. [cited 2013 28.01.]; Available from: http://www.infostat.sk/slovakpopin/.
20. Ripley, B. R. Data Import/Export. 2012: R Core Team
21. NCZI. Healthcare Almanacs. Edition of Analytic Publications [cited 2013 31.01.]; Available from: http://www.nczisk.sk/Publikacie/Pages/Edicia-analytickych-publikacii.aspx.
22. NCZI. National Health Registers. [cited 2013 02.02.]; Available from: http://www.nczisk.sk/ Registre/Narodne-zdravotne-registre/Pages/default.aspx.
23. Murray, C. J., Lopez, A. D., Wibulpolprasert, S. Monitoring global health: time for new solutions. BMJ, 2004. 329(7474): p. 1096 – 1100.
24. Norwegian Social Science Data Services. MacroDataGuide. 2013 [cited 2013 16.02.]; Available from: http://www.nsd.uib.no/macrodataguide/.
25. Business dictionary. Web portal. [cited 2013 16.02.]; Available from: http://www.businessdictionary.com/definition/portal.htm
26. WHO EURO. European Health for All Database. [cited 2013 16.02.]; Available from: http:// data.euro.who.int/hfadb/.

27. WHO. WHO Mortality Database. 2012; Available from: http://www.who.int/whosis/mort/do-wnburden/en/index.html.
28. WHO. Indicator definitions and metadata. 2012 16.12.2012]; Available from: http://www.who. int/whosis/indicators/en/.
29. EUROSTAT. EUROSTAT-DataBASE. 2013 [cited 2013 19.02.]; Available from: http://epp. Eurostat.ec.europa.eu/portal/page/portal/statistics/search_database.
30. Statistics Sweden. PC-Axis web site. [cited 2013 11.03.]; Available from: http://www.scb.se/Pa-ges/StandardNoLeftMeny____314045.aspx.
31. IBM. IBM SPSS Statistics [cited 2013 11.03.]; Available from: http://www-01.ibm.com/soft-ware/analytics/spss/
32. ISI. International Statistical Institute. [cited 2013 20.08.]; Available from: http://www.isi-web. org.

Chapter 6

Data standardization

Chapter objectives

Mortality must be often compared in two or more populations, let´s say in two countries, districts or towns. We know that one of them is much younger with more children and young people, while the other one has fewer kids and many old people. Logically, mortality is higher in the country with older people. Gross mortality reflects the differences caused by the mortality and those in the population age structure. Method of the problem solving is called standardization and it is frequently used in comparison of health indicators. Mortality comparison was a basic motivation of the standardization implementation. Despite of possibility to standardize in the terms of many variables that could influence its development, age is most frequently used at mortality standardization because of its impact on illness rate and mortality.

History in brief

In the half of the 19th century, British public health experts started to recognize that rough vital statistics are not suitable for comparison of public health indicators in groups with too different age structure. Discussion was held at the time on the mortality index development that wouldn´t be influenced by age differences. In the lecture for the Statistic Society in London, Mr. Edwin Chadwick, known as one of the first public health reformist in England, proposed application of the indicator "death age median"

Age group	European population	WHO population
0-4	8.00	8.86
5-9	7.00	8.69
10-14	7.00	8.60
15-19	7.00	8.47
20-24	7.00	8.22
25-29	7.00	7.93
30-34	7.00	7.61
35-39	7.00	7.15
40-44	7.00	6.59
45-49	7.00	6.04
50-54	7.00	5.37
55-59	6.00	4.55
60-64	5.00	3.72
65-69	4.00	2.96
70-74	3.00	2.21
75-79	2.00	1.52
80-84	1.00	0.91
85+	1.00	0.63
Total	**100.00**	**100.00**

Table No. 1 Standard populations for EU population and WHO recommended for all world countries [3]

as a tool of health condition comparison in various London parts. To his opinion, the index represented a real view on age-specific death risk. [1] Neison who dealt with insurance mathematics disagreed with his calculations and proved that he was mistaken. He also presented the solution called "direct and indirect standardization" known under this name to date, and the term "standard population".

Neison´s method of direct standardization was used already in 1883 in the Population Report in England with census results from 1881 in Wales and England, used as a standard. The standard was calculated per each new census in the following reports, i.e. once per 10 years. Such frequent changes of the standard were laborious, since it was necessary to make new historical calculations in order to evaluate current trends. Finally, census dated 1901 was adopted as a standard solution in England and Wales. It had remained unchanged even if new census data were available. With the aim to simplify comparison to the mortality in England and Wales, USA adopted the British standard in 1901. This system had been used till the early 40s when it was decided that the difference between USA population at the time and British population in 1901 was so significant that it was found a sufficient reason for the standard modification. [2] In 1892, Ogle recommended establishing international comparison standard.

Since then, various standards had been proposed but no one was accepted generally. The Union against Cancer presented three standards in 1965. WHO had used all of them until the proposal and adopted of completely new standard. [3] Table No. 1 contains standard values for European and worldwide population. Studying the population health, we will experience situations that can distort the relation between variables or bear a systematic error. In both cases, application of one of standardization methods represents one of the solutions. We will describe the situations in detail in the following text.

Distortion and bias

Sometimes we can observe causal relation between the variables that is, in fact, not present. Variables distorting relations between further variables being subject of our investigation are called "*confounders"* [4]. To be a confounder, the variable must correlate with disease or condition we are interested in and the studied risk factor. [5] The results are biased until the bias is removed. Let´s talk in detail about these processes we can often encounter in professional literature but their full content is not always fully understood.

Bias, distortion

In the statistics, a process is considered distorted if not all possible outputs are equally probable; in such case we say that it is bias from most likely output. When we throw a die, we automatically presume the same likelihood of each die side falling if we throw many times. Die throwing is unfair if not all reached scores are equally possible. If we report more score 3 or 5 after many throws and very few other values, we can expect that the die was biased or falsified. Epidemiological dictionary defines bias together with its causes: *Biased results or conclusions from the reality or processes that lead to such bias. Any trend within collection, analysis, interpretation, publication or verification of data can lead to the conclusions and systematically differ from the reality. The following are some of the processes with possible bias occurrence: 1. Biased measurement from real value. 2. Variations of statistical summary measures (diameter, proportion, association extent, etc.) from the real value as a result of measurement bias, other data collection shortages or deficiencies in proposed study or analysis. 3. Deviation of conclusions from real situation as a result of deficiencies in the draft study, data collection or analysis, or interpretation of results. 4. Process tendency (draft study, data collection, analysis, interpretation, evaluation* *or publication) to bring results or conclusions that deviate from real situation.*

5. Bias leading to conscious or unconscious selection of study processes that differ from real situation in certain direction or lead to unilateral interpretation of results. The term bias, distortion doesn´t necessarily have to indicate bias or other subjective factor, e.g. wish of particular result by the experimenter. This is the difference from colloquial use when bias refers to certain party opinion (preference). [6]

Dictionary contains large amount of sources and causes of bias. Classical thesis of Sackett deals in detail with bias in the area of clinical research. [7] For example, Rodriguez and Llorca describe bias in the area of epidemiology [8]. History of epidemiological bias identification is very interesting and well-illustrating the development of human thinking. [9] For purposes of this publication, we will discuss a few of them in order to know those important ones when working with data from routinely collected statistics.

Detection distortion

Distortion by detection is caused by bias in the methods of searching, diagnostics or case verification in epidemiological study. Example – diagnosis verification with laboratory tests in and outside the hospital caused by different tests made. It can refer to tuberculosis patients headcount based on field screening test when hospitalized patients who passed laboratory tests were counted in.

Availability distortion

Availability of health and healthcare services is usually the same within a territory. This method can be influenced for example by healthcare staff interest in certain cases (popularity distortion) when the facility chooses certain patients for example for pharmaceutical testing. Sometimes, availability has impact on the prestige of certain facilities or physicians/ nurses (centripetal distortion). Another case refers to situation when serious cases are moved to higher healthcare level in the terms of hierarchy (filter distortion), etc.

Report distortion

Distortion upon status announcement can result for example from the efforts to satisfy a researcher and respond in the way that is presumed suitable or awaited (servility distortion). Sensitive questions that can make a requester embarrassed or hurt him/her socially, can be rejected (unacceptable diseases/ exposures) – for example alcohol intake question or intimate behaviors.

Classification distortion

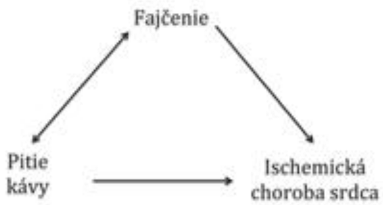
Distortion by classification often results from ambiguous interpretation of classification criterions. For example, if inexperienced physician or nurse assesses the patient´s condition, for example consciousness level or injury degree. Their assessment can differ from that of report made by experienced colleague. The situation occurs also upon absent/ not observed agreement on classification rules. For example, heart failure (ICD I50) is most frequently stated as death cause instead of its most frequent cause, which is hypertension (ICD I11), despite of specifically noted in the classification. It results in undervalued hypertension occurrence that caused the death.

Confounding – confusion

Dictionary of epidemiology defines confounding (Latin *confundere*) as: *1. Condition of not separated effects of two processes. Mixing of virtual effect of exposure on the risks resulting from connection with other factors having potential to affect the result. 2. Relation between the effects of two or more causal factors observed in the data file in such way that it is impossible to logically separate contribution to the effect of each one causal factor. 3. Situation when the rate of exposure effect on the risk is distorted by the exposure relation with other factor (s), affecting examined result.*

[6] To illustrate the concept, the following is an example published as a part of review article about measure standardization in epidemiology. [10]

In the study dedicated to coffee drinking as a risk factor of myocardial ischemia development, the fact if the subject was/ wasn´t a smoker had to be considered as well. Tobacco smoking is a generally recognized factor having significant influence on the disease development. It is also known that tobacco and coffee are usually consumed simultaneously but we cannot say that smoking is a coffee drinking consequence. Thus, we cannot consider smoking a variable that would be positioned between coffee and myocardial ischemia. Simple picture No. 1 shows the relation between the three variables.



Picture No. 1 Correlation between coffee drinking, smoking and myocardial ischemia pursuant to [10]

Connecting lines indicate that if we examine effects of coffee drinking on the myocardial ischemia development, we have to consider mutual effect of coffee and tobacco on the myocardial ischemia onset only in one direction. If we avoid the tobacco role, we would make a mistake. Such situation is called confusion and the factor is confounding.

Confounding variable, confounder

Confounding variable can lead to false result if not considered. It must be associated with factor being subject to the investigation; otherwise the result would be distorted. [6] Such factor (confounding variable) must meet the following criterions: defined risk factor in relation to examined result, an exposure-related factor but not result of the exposure, it shouldn´t be posed in-between two primarily studied variables (exposure and effect).[11]

Confounding is not a program error but a phenomena identified through investigation that should be taken in account. Distortion results from the study process error but confounding is a valid finding describing the nature of relation between a few factors and the disease risk. However, inability to accept confounding variable during interpretation of investigation results refers to the study implementation error that can affect its conclusions. [11]

Considering the efforts to eliminate the effect of confounding variables as much as possible and remove or reduce their impact, we use standardization of measures as a basic method. In practice, we often standardize age, mainly when comparing measures or indicators derived from mortality. Other confounding variables are standardized less frequently. Similar to any other measure resulting from the combination of more variables, standardized variable can cover large differences between the groups that could be important for explanation of changed measures as a result of or in relation to the variable to be modified. Thus, always when possible, it is important to analyze a measure standardization together with those non-standardized.

Standardization methods

In fact, two methods are used for removal of the differences in demographical structure of two populations: direct and indirect method. If used with the same input data, they should give the same result. If we obtain two different results, we should search for the cause through deeper examination of input data properties. Regardless the applied method, we have to determine the standard population we want to use. It can be one of the compared populations or a completely different or theoretically created population.

Direct standardization method

Direct standardization method shall be demonstrated on the example of mortality comparison in two Bratislava municipalities: Bratislava I, the Old City with few young people living, and Petržalka, (Bratislava V) with many young people living.

	Bratislava I (a)			Bratislava V (b)			B V/BI
Age group (i)	Total deaths x_{ia}	Mortality /1000 p_{ia}	Population n_{ia}	Total deaths x_{ib}	Mortality /1000 p_{ib}	Population n_{ib}	Population n_{ib}/n_{ia}
0-- 14	6	1.3	4 636	5	0.4	12 617	2.7
15-- 59	63	2.7	23 595	241	2.9	83 491	3.5
60+	490	38.9	12 597	283	13.2	21 464	1.7
In total	559	13.7	40 828	529	4.5	117 572	2.9

Table No. 2 Total deaths and population in Bratislava I and Bratislava IV in 2009, both genders included. Source: SOURCE: STATISTICAL OFFICE

Both populations don´t show big difference in the count of the deaths. However, if we look at the ratio between population in Bratislava I- Old City and Bratislava V-Petržalka, we can see an interesting difference (Table No. 2). Population in the Old City has fewer young people than Petržalka. We standardize with reference to a common population similar to both mentioned. In fact, it is equal what population we intend to standardize. For illustration, let´s choose population resulting from the sum of all citizens living in Bratislava Old City and Petržalka in each age group. Using mortality per 1000 citizens in each community, we recalculate total deaths per the new population. Summarizing the results, we obtain resulting mortality per each community. (Table No. 3).

Age (i)	Population BA I + BA V $P_i=n_{ia}+n_{ib}$	Bratislava I			Bratislava V		
		Mortality /1000 p_{ia}	Expected Count of deaths e_{ia}	Standar-dized mortality sdr_{ia}	Mortality /1000 p_{ib}	Expected Count of deaths e_{ib}	Standar-dized mortality sdr_{ib}
0-14	17 253	1.3	22.3		0.4	6.8	
15-59	107 086	2.7	285.9		2.9	309.1	
60+	34 061	38.9	1 324.9		13.2	449.1	
In total	158 400		1 633.2	0.01031	4.499	712.7	0.00450

Table No. 3 Calculation of standardized mortality through direct standardization method

Let´s explain the calculation step by step. At first, we should say that we chose only 3 age groups to simplify the calculation. They are marked *i*. Community Bratislava I is marked with *a*, Bratislava V with *b*. Total deaths in particular age groups are marked as x_i, with two different variables - x_{ia} for the 1st community and x_{ib} for the 2nd community. Analogically, populations shall be marked as n_{ia} and n_{ib}. Gross mortality rate in particular age group and community shall be marked as p_{ia} and p_{ib}. It is calculated then as a simple proportion of total deaths in particular age group to the average population in particular age group and community:

Community *a*: $p_{ia} = \frac{x_{ia}}{n_{ia}} 1000,$

Community *b*: $p_{ib} = \frac{x_{ib}}{n_{ib}} 1000.$

Summing up the population in both municipalities according to the age groups, we obtained common or standard population *Pi*, to which we recalculate the death counts from particular municipalities. In this way, we obtain the value of expected count of deaths in particular age groups and municipalities e_{ia} a e_{ib}:

Community *a*: $e_{ia} = \frac{p_{ia}}{P_i} 1000,$

Community *b*: $e_{ib} = \frac{p_{ib}}{P_i} 1000.$

Standardized age-based mortality marked as sdr_i will be calculated similar to the proportion of expected mortality and average population in the community:

$$sdr_{ia} = \frac{e_{ia}}{n_{ia}} 1000,$$

$$sdr_b = \frac{e_{ib}}{n_{ib}} 1000.$$

Interpretation of the result refers to comparison of standardized age-based mortality applicable to both municipalities. We can see that it is much higher in Bratislava I, Old City than in Petržalka, Bratislava V. Resulting from the relative risk concept, we can interpret the mortality rates as probability of dying and their ratio as a relative death risk:

$$RR = \frac{sdr_a}{sdr_b} = \frac{0{,}01}{0{,}0045} = 2{,}29.$$

Thus, we can state that relative death risk is 2.3 times higher for Bratislava Old City inhabitant than for Petržalka inhabitant.

It´s not always suitable to use combination of participating populations as a standard. Population defined by WHO has been used in the health and disease studies. It is much easier and represents currently a global standard, used practically always in case of international comparisons. Investigating mortality or other health-related indicator through comparison of two or more geographical or administrative wholes, population of higher age is usually used, for example higher territorial whole (VÚC) in case of investigated districts or the whole country. Selection of standard population wouldn´t significantly influence the result.

Epitools

To simplify the calculation of standardized measures through direct and indirect standardization, project {R} offer the library of epidemiological methods, called *epitools*. [12] We will demonstrate the use of function *ageadjust. direct(count, pop, rate = NULL, stdpop, conf.level = 0.95)* from the library on the previous example; calculating standardized mortality including the uncertainty index. Procedure of data preparation in the environment Rexcel and obtained results are apparent on the following picture: (Picture No. 2).

```
# data were inserted in variables, removing them from the table in RExcel with Put R Var
# total deaths in community A, Bratislava I
> countA
[1]  6 63 490
# total deaths in community B, Bratislava V
> countB
[1] 5 241 283
# inhabitants´ headcount in community A, Bratislava I
> popA
[1] 4636 23595 12597
# inhabitants´ headcount in community B, Bratislava V
> popB
[1] 12617 83491 21464
# standard population
> stdpop
[1] 17253 107086 34061
# calling function for community A result
# variable Gross.rate contains value of gross mortality rate, variable adj.rate contains standardized
mortality rate, variable lci states bottom uncertainty interval and uci refers to top
uncertainty interval
```

> *ageadjust.direct(pocetA, popA, rate = NULL, stdpop, conf.level = 0.95)*
Gross.rate adj.rate lci uci
0.013691584 0.010310384 0.009456982 0.011230608
> *ageadjust.direct(pocetB, popB, rate = NULL, stdpop, conf.level = 0.95)*
Gross.rate adj.rate lci uci
0.004499371 0.004829766 0.004424762 0.005262736

Community	Gross rate	Standardized	Bottom interval	Top interval
Bratislava I	0,014	0,010	0,009	0,011
Bratislava V	0,004	0,005	0,004	0,005

Picture No. 2 Data preparation process and direct standardization results

Now look what happen if we apply other standard population. We will start with the one used by WHO. [3]

#we entered particular age groups in the variable *stdWHO*
> *stdWHO*
[1] 26.15 61.93 11.95
#Calling function *ageadjust.direct()* we calculate standardization values
> *ageadjust.direct(pocetB, popB, rate = NULL, stdWHO, conf.level = 0.95)* Gross.rate adj.rate lci uci
0.004499371 0.003465815 0.003167792 0.003800535

> *ageadjust.direct(pocetA, popA, rate = NULL, stdWHO, conf.level = 0.95)*
Gross.rate adj.rate lci uci
0.013691584 0.006638342 0.006013983 0.007359279

Community	Gross rate	Standardized	Bottom interval	Top interval
Bratislava I	0,014	0,007	0,006	0,007
Bratislava V	0,004	0,003	0,003	0,004

Picture No. 3 Standardization of mortality in two municipalities of Bratislava, using population standard of WHO

You can see from the result (Picture No. 3) that application of WHO standard population didn´t significantly change the standardization result.

Indirect standardization method

Indirect standardization is chosen in case of insufficient data for direct method and it is less frequently used. Indirect method results from standard variety of age-specific mortality that is correlated to total mortality of population in question.

Age-specific mortality is obtained from the reference population. In this way, we obtain "expected" mortality in the observed population, and recalculate the age-specific mortality of standard population to the monitored population. Most frequently stated ratio between observed and expected mortality is also called "Standardized Mortality Ratio" or SMR. [13]

Let´s continue with the process of indirect standardization on the textbook example taken from [14]. Input data and calculation result are in two tables on the picture No. 4. Expected mortality in each of the studied populations was obtained through multiplication of monitored age-specific population median with standard mortality. The result was recalculated per 1000 inhabitants.

$$e_{ia} = \frac{n_{ia} p_i}{1000}$$

Summary indicator of expected mortality ratio to total population was obtained through division of expected mortality in the whole monitored population A and B and division of total population median in each of the municipalities:

$$A_a = \frac{e_a}{n_a} 1000 \qquad B_a = \frac{e_b}{n_b} 1000$$

Then we performed indirect standardization; we multiplied gross mortality rate with mortality of the whole population and divided the result by the ratio of exported mortality and total population:

$$imdr_a = \frac{p^* mdr}{A_a} \qquad imdr_b = \frac{p^* mdr}{A_b}$$

Index of standardized mortality represents a simple ratio of indirectly standardized mortality rates:

$$imdr = \frac{imdr_a}{imdr_b}$$

Index *imdr* (Indirect Method Death Rate) states that indirectly standardized mortality in population A is lower than in population B by approx. 10%. If they were equal, then *imdr* would refer to 1. Now we will make the calculation easier, using function from EPITools, which will calculate particular mortality rates and introduce the uncertainty internals. We will use data from the example of two Bratislava municipalities comparison in the calculation but with a small difference; namely we will draw standard population and corresponding death numbers from the statistics of the whole Slovakia in 2010 (Picture No. 5).

	Standard population			Population in community	
Age group	Population	Deaths	Mortality /1000	A	B
0-- 14	35 000	330	9.43	10 000	25 000
15-- 59	30 000	345	11.50	15 000	15 000
60+	35 000	535	15.29	25 000	10 000
In total	100 000	1 210	12.10	50 000	50 000
Gross mortality rate/1000	***mdr***	**12.40**	**11.80**		

Standard mortality/1000 P_i	Population ni_a	Expected deaths e_{ia}	Populatio n n_{ib}	Expected deaths e_{ib}
9.43	10 000	94.29	25 000	235.71
11.50	15 000	172.50	15 000	172.50
15.29	25 000	382.14	10 000	152.86
12.10	50 000	648.93	50 000	561.07
Ratio of expected deaths to total population A_a		**12.98**	A_b	**11.22**
Indirectly standardized mortality $imdr_a$		**11.56**	$imdr_b$	**12.72**
Index of standardized mortality $imdr$		**0.91**		

Picture No. 4 Example of standardized mortality index calculated through indirect standardization method

	Bratislava I A		Bratislava V B		Slovakia	
Age group	Total deaths *countA*	Population *popA*	Total deaths *countB*	Population *popB*	Total deaths *stdcount*	Population *stdpop*
0-14	6	4 636	5	12 617	505	831 246
15-59	63	23 595	241	83 491	10 133	3 636 173
60+	490	12 597	283	21 464	42 807	963 605
Total	559	40 828	529	117 572	53 445	5 431 024

Picture No. 5 Input data of indirect standardization example. Names of variables – arguments of called function – as stated in italics.

Calling the function *ageadjust.indirect(count, pop, stdcount, stdpop, stdrate=NULL, conf.level = 0.95) from* EPITools, we will standardize mortality rate at first for Bratislava I (marked with A) and then for Bratislava V (marked with B).

```
> ageadjust.indirect(countA, popA, stdcount, stdpop, stdrate=NULL, conf.level = 0.95)
$sir
  observed      exp      sir      lci      uci
559.0000000 628.1758184   0.8898783  0.8190843 0.9667910
$rate
 Gross.rate  adj.rate      lci      uci
0.013691584 0.008757012 0.008060351 0.009513886
```

SMR populations A

Indirectly standardized mortality of population A

```
> ageadjust.indirect(countB, popB, stdcount, stdpop, stdrate=NULL,
conf.level = 0.95) $sir
  observed      exp      sir      lci      uci
 529.0000000 1193.8437692 0.4431066 0.4069110 0.4825218
$rate
 Gross.rate  adj.rate    lci    uci
0.004499371 0.004360472 0.004004283 0.004748345
```

SMR populations B

Indirectly standardized mortality of population B

Community	Gross rate	Standardized	Bottom interval	Top interval
Bratislava I	0,014	0,009	0,008	0,009
Bratislava V	0,004	0,004	0,004	0,005

Picture No. 6 Standardization of mortality according to age in two communities using indirect standardization method.

You can see similar numbers obtained from the indirect standardization method results, applied to the data of two communities similar to those in case of direct standardization but using other standard population. (Picture No. 6)

Application of standardized mortality can be disputable. Any aggregate measure can overlap the factors that could have major effects on the public health. On the example in case of age- standardization, one could oversee age-specific differences in the risk in the terms of time or location. Such situation can occur in case of virtually increasing occurrence of new cases of a disease as a result of birth cohort effect (younger people could face higher disease risk than the older ones). Age-standardized measures provide useful information despite of the risk, mainly in cases of rare diseases and wide variability of specific measures.

Life expectancy

People have been asking about the life expectancy for ages. This subject matter acquired the experts´ attention only lately, thanks to insurance industry foundation. Everything has been insured; from ships to houses and factories. When life insurance appeared, not only detective stories were written but it was necessary to calculate whom it is worth to insure and whom not. Let´s take an example: an old rich man decides to conclude insurance contract for premature death with the clause that if he dies before the specified date, insurance premium shall be paid to the widow or kids. Insurance company is a business subject and as such, it performs activities to reach profit. Thus, insurance payment should be higher than actually paid up insurance premium against unexpected situation/ accident. If it was insurance of a single person, the insurance company would face risk that the insured person dies before the agreed date and the committed insurance premium would exceed the amount paid to date. However, insurance companies apply rules that eliminate such cases. They deal with the issue – how many years could an individual live, taking in account particular age, gender, etc. Based on estimated life expectancy, the company is able to calculate insurance premium and payments. It is still possible that the insured person dies earlier but there are not many such cases if the life expectancy is correctly determined, and insurance companies used to be highly profitable.

Actuarial mathematics

How to estimate whether somebody lives up to certain age? If we have sufficiently large group of people available and know the death likelihood in particular age groups, we can predict headcount of those who will survive with sufficient certainty. Such predictive mathematics is called *actuarial*. Based on such calculation, we can calculate average whole years of life, commonly named as *life expectancy*. We will calculate the survival likelihood according to death probability in particular age category. These calculations were made with older people´s data, sometimes with the difference of a few generations. In other words, if we want to investigate likelihood of living up to 60 years of age for currently born child, we will do it on basis of grandparents´ data (grandparents born at the beginning of the 50s of the previous century). Thus, we project facts in the life expectancy that influenced the life of people in this age category and these could significantly differ from the circumstances that the newborn will face in his/her life and that cannot be foreseen. Therefore, actuarial statistics provides historical image of the presence and its interpretation is based on precondition that estimations (for example for the newborn) will be correct only if equal behavior of certain age structures is considered. We know that such precondition is always hypothetical and bears bias. Neither can we say that current newborn will live up to the age as presumed according to the life expectancy.

Life expectancy has an advantage, namely it represents summary mortality and doesn´t require standard population application. Thus, process of its calculation can be considered age-related standardization and mutual comparison of populations. It can be also easily compared within territorial wholes and countries and is generally understandable. Of course, its interpretation value of the population health condition is limited, similar to the data about age-standardized mortality. Simple definition of life expectancy at birth, expectation of life at birth says that it is *number of years that will be lived up by a born person in average, provided that mortality conditions don´t change*. [15]

Life expectancy (LE) is the most frequently used global characteristics of mortality assessment. Considering the use of this measure, we will present its calculation and the way of related considerations. We will also describe proper interpretation of results and their application to considerations of the factors taken in account during interpretations of the results. We will also present our considerations of the life and health quality when comparing various population groups.

History

Mortality tables have long history. John Graunt published the book *Natural and Political Observations upon the Bills of Mortality* in 1662 based on the Bills of Mortality collected by him. [16] Similarly, Edmund Halley (1656 – 1742), recognized comet discoverer, published the tables from the German town Breslau (current Wroclaw) in 1693, compiled by honorable Caspar Neumann (1648 – 17I5). The tables included 5-year period of years 1687 – 1691. From the tables, Halley derived the life expectancy and published the calculation in the essay *An estimate of the degrees of the mortality of mankind, drawn from the curious tables of the birth and funerals at the city of Breslaw, with an attempt to ascertain the price of annuities upon lives* (annuity[21]). The importance of mortality tables as a calculation tool found smaller response by the statists than by actuarial mathematicians. [17]

Mortality tables

Mortality tables or life tables describe the process of population dying. The principle of the construction is based on the determination of a person death likelihood based on the age in certain period and on decrease of the table headcount of people living on basis of the likelihood.

---

21 Annuity (annuity payment) is payment of credit including credit principal payment and interest payment.

Mortality tables are calculated separately according to genders for the age within 0 – 104 years. [15] They represent a tool for identification of mortality conditions in particular population [18], providing the review of death likelihood and life expectancy based on age group and gender. Mortality tables are sometimes also called Life Tables and represent a model, instead of real status, of death behavior of an enclosed population where birth is not taken in account and death of an individual is the only possible life termination. Thus a model or abstraction of the tables are based on death and doesn´t count with migration, newborns, etc.

Generation (cohort) mortality tables

If the table captures the life course record at particular population of concurrently born individuals from the birth till the death of the last one of them, we speak of generation *mortality table*. Compilation of such table is complicated since the population must be monitored during its entire life. Of course, losses could occur during such human cohort monitoring, for example through migration or lost follow-up. Accordingly, the tables are mainly used during monitoring of animals, insects or bacteria because of shorter life than at humans. Lately they have been used also in epidemiological research.

Common (cross-sectional) mortality tables

These tables describe mortality rate of certain population within the time period. Based on mortality rates according to particular age groups, the picture of life of hypothetical population of concurrently born individuals is constructed. They are divided in complete and abridged life tables according to determined age interval. Complete life tables are calculated by the functions on annual basis. Abridged tables use the interval longer than 1 year except the 1st year of life. [14] They represent the basis of life expectancy calculation.

Mortality table calculation

Mortality table calculation input represents the count of dead persons in particular age and the count of live people in the given age (median), all within the monitored year. Calculation procedure is stated on INFOstat. [18]

Abridged life tables are derived from detail mortality tables and calculated per age groups. As we stated in the definition, age interval width x is bigger than 1, except the age group up to the 1st year of life.

Life expectancy and abridged mortality tables calculation is rather simple but requires much effort to observe all calculation principles. It is possible also in Excel format. These tables are usually not calculated anymore but overtaken from the professional demographic sources; or a special SW is used for calculation. We will explain the calculation methodic through simple words, using mortality tables 2010 prepared by INFOstat and freely available on its website.[22] The count of dead persons in particular age and the count of live people in the given age (median), all within the monitored year, represent the only input data required for preparation of the mortality table. Picture No. 7 presents such a mortality table (or its part within 0 - 10 years, within 90 -100 and more years) applicable to the Slovak population in 2010. The following is a closer look:

Podrobné úmrtnostné tabuľky
Územie: Slovenská republika

Obdobie:2012
spolu

Vek	Zomreli	Žijúci	q_x	l_x	d_x	L_x	T_x	e_x
0	321	58213	0,005499	100000	550	99505	7606521	76,07
1	31	59402	0,000522	99450	52	99424	7507016	75,49
2	15	58813	0,000255	99398	25	99386	7407591	74,52
3	8	58441	0,000137	99373	14	99366	7308206	73,54
4	13	55790	0,000146	99359	14	99352	7208840	72,55
5	6	54236	0,000093	99345	9	99340	7109488	71,56
6	6	54271	0,000102	99336	10	99330	7010148	70,57
7	3	54268	0,000110	99325	11	99320	6910817	69,58
8	11	52849	0,000170	99314	17	99306	6811497	68,59
9	8	51294	0,000162	99298	16	99289	6712191	67,60
10	10	51218	0,000151	99281	15	99274	6612902	66,61
90	1125	5209	0,222650	12550	2794	11153	37844	3,02
91	891	3700	0,248887	9756	2428	8542	26691	2,74
92	703	2561	0,277747	7328	2035	6310	18149	2,48
93	368	1459	0,309326	5293	1637	4474	11839	2,24
94	202	744	0,343678	3655	1256	3027	7365	2,01
95	183	581	0,380795	2399	914	1942	4338	1,81
96	140	501	0,420592	1486	625	1173	2395	1,61
97	161	437	0,462890	861	398	662	1222	1,42
98	96	336	0,507397	462	235	345	561	1,21
99	88	249	0,553697	228	126	165	216	0,95
100+	109	383	0,601238	102	102	71	71	0,70
Spolu	**52 437**	**5 407 580**						

Picture No. 7 A part of detail mortality tables, Slovakia 2012. Source: Slovak Statistical Office

The first column contains the age of death; the second column states headcount of the dead and the third one contains headcount of the live. The following column marked q_x represents the likelihood of dying in particular age group. It is an estimated likelihood that an individual of the age equal to x will die in particular year. Estimation is done on basis of age-specific mortality rate according to rather

22 http://portal.statistics.sk/files/Sekcie/sek_600/Demografia/Obyvatelstvo/tabulky/Ut/2012/ut-2012.pdf

complicated formula that can be found in the references.[14, 18] The next column starts with population model considering 100 hundred newborns (table headcount of the surviving for the age 0) that will die gradually according to the death likelihood in particular age. In this way, we will get gradually reduced headcount of the surviving in the artificial population, marked l_x, - table headcount of the surviving. The following column L_x contains the table headcount of the dead. Notice that the difference of two following counts of the surviving in the table refers to the table count of the dead in the first of the years in the difference. The column marked d_x refers to the average table count of the surviving (so called stationary population). Its calculation is also a bit complicated: each member of the cohort who lives up to the end of the age interval will add one year to the value L_x, while each member who dies will add only the part that was lived up within the interval. To make it simpler, it is often presumed that the dead died in the middle of the interval. Last but one column of the table marked T_x contains the years of life ahead of the table generation (not an individual) of particular age. It is calculated per each age as the count of the remaining years of life, i.e. $T_x = L_x + L_{x+1} ... L_{100}$. And finally, value of potential living up to particular age is calculated as a ratio of the average table count of the live and the table count of the surviving.

$$e_x = \frac{T_x}{l_x} .$$

Skrátené úmrtnostné tabuľky
Územie: Slovenská republika

Obdobie:2012
spolu

Vek	Zomrelí	Žijúci	q_x	l_x	d_x	L_x	T_x	e_x
0	321	58213	0,005499	100000	550	99505	7606521	76,07
1-4	67	232445	0,000288	99450	105	99397	7499265	75,41
5-9	34	266918	0,000127	99345	63	99313	7101675	71,49
10-14	43	273999	0,000157	99281	80	99242	6605110	66,53
15-19	128	324444	0,000394	99202	190	99107	6108902	61,58
20-24	220	392065	0,000561	99012	284	98870	5613368	56,69
25-29	278	431779	0,000644	98728	310	98573	5119018	51,85
30-34	309	455289	0,000678	98418	343	98247	4626151	47,00
35-39	567	444111	0,001276	98075	613	97769	4134918	42,16
40-44	772	364976	0,002113	97462	1037	96944	3646074	37,41
45-49	1334	369345	0,003605	96426	1733	95559	3161353	32,79
50-54	2309	375688	0,006127	94693	2841	93272	2683558	28,34
55-59	3596	387971	0,009226	91851	4170	89766	2217199	24,14
60-64	4723	329898	0,014215	87681	6142	84610	1768368	20,17
65-69	4698	227726	0,020419	81539	8048	77515	1345317	16,50
70-74	5542	179183	0,030456	73491	10651	68166	957742	13,03
75-79	6978	134370	0,050606	62840	14518	55581	616914	9,82
80-84	8888	95577	0,088800	48322	18190	39227	339008	7,02
85+	11630	63587	0,167148	30132	30132	142873	142873	4,74
Spolu	**52 437**	**5 407 580**						

Picture No. 8 Abridged life table Slovakia 2012. Source: Slovak Statistical Office

Calculation of the abridged mortality tables is derived from the complete tables. Resulting table for Slovakia in 2012 is on the picture No. 8.

Let´s try to interpret the table. We will refer to the Mésáros [19] list of possible use of the mortality tables (Picture No. 9).

l_y/l_0	Likelihood of surviving from birth l_0 (up to particular age y (L_y)
l_y/l_x	Likelihood of surviving from age x to age y
$1 - l_y/l_x$	Likelihood of dying between ages x and y
$L_x - l_y/l_0$	Likelihood of a newborn dying between ages x and y
$T_x - T_y/l_0$	Number of years survived by a newborn between ages x and y
Normal length of life	modus of the table deads d_x - age in which people die most frequently (except age 0)
probable Length of life	median of the table surviving l_x, age lived up by 50% of the born, considering particular mortality rate, resp. the age x to which applies: $l_x = 1/2 l_0$
Lost years of The dead´s life	If a person lives up to age x, he will probably face e_x more years. If he dies in age x, he lost e_x years. Not everybody dies in the same age, thus we calculate lost years as the mean of two consecutive years: $v_x = \frac{e_x + e_{x+1}}{2}$
Lost years of life For the whole population	$V = \sum_{x=0}^{\omega} D_x^{muži} * v_x^{muži} + \sum_{x=0}^{\omega} D_x^{ženy} * v_x^{ženy}$

Picture No. 9 Possible interpretations of mortality tables. Mésáros [19]

At first it contains probability of surviving from the birth up to particular age. If we look at the detail table of survivors for Slovakia 2010, the likelihood of a Slovak person to live up to 50 years of age refers to 0.94, i.e. 94 % of newborns would live up to 50 years of age.

Table No.4 shows that approx. 92 % of men born in 2010 will live up to 50th birthday while there will be more such women. This difference increases in the next life decades up to the 90th year of age that will be celebrated by more than 50% more women than men. This aspect is rarely used in the expert literature; however it is a perfect basis for the consideration of the examined population condition.

To illustrate another mentioned use, we will do calculation based on the previous data and try to estimate the likelihood of living up to higher life decades by current 50 and 70 years old people (Table No. 5).

Age of survival y	l_y		Surviving likelihood	
	Men	**Women**	**Men**	**Women**
50	92353	96664	92%	97%
60	81778	92403	82%	92%
70	61644	82522	62%	83%
80	32537	57527	33%	58%
90	5875	14022	6%	14%
100	43	35	0,04%	0,03%

Table No. 4 Likelihood to live up to certain age based on detail mortality table for Slovakia, 2010. l_0 = 100 000. Source: POPIN

Initial age x	Survival age y	l_x		l_y		Surviving likelihood	
		Men	**Women**	**Men**	**Women**	**Men**	**Women**
50	60	92353	96664	81778	92403	89%	96%
	70			61644	82522	67%	85%
	80			32537	57527	35%	60%
	90			5875	14022	6%	15%
	100			43	35	0,05%	0,04%
70	80	61644	82522	32537	57527	53%	70%
	90			5875	14022	10%	17%
	100			43	35	0,0007	0,0004

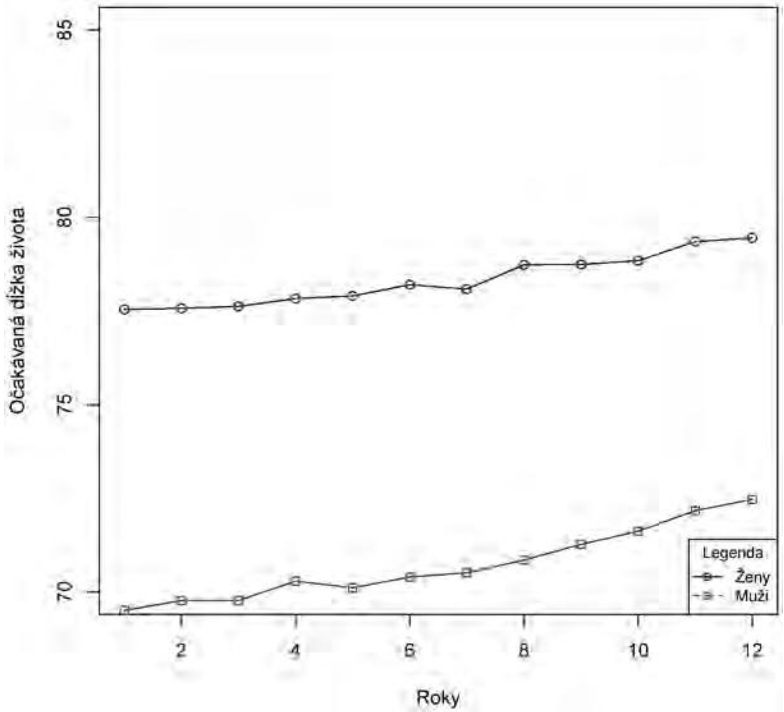
Table No. 5 Likelihood of living up from age x to age y based on mortality table for Slovakia 2010
Source: POPIN

It is apparent that the men who lived up to 50 years of age have smaller than 90% likelihood that they live up to the following decade. They will live up to 70 years of age with likelihood smaller than 70 %, while for women of the same age is this likelihood by 8 % higher.

PROS and CONS

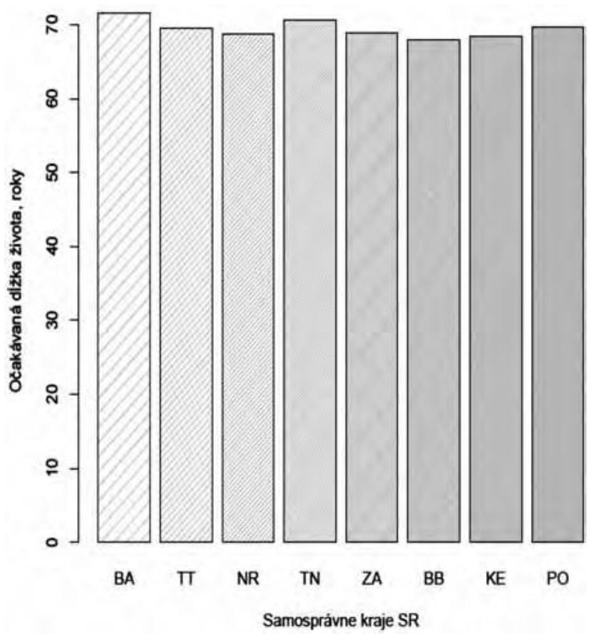
Despite of relatively complicated calculation of the life expectancy, it brings intuitively easily understandable result. This feature is, however, associated with the risk of incorrect interpretation that doesn´t take in account retrospective basis of the calculation or the likelihood of static mortality trend. Comparison of the life expectancy between two or three populations can neglect the age-based standardization and it is major advantage to the mortality. Nevertheless, this value is strongly affected by early-age mortality. Life expectancy at birth in the countries with high infant age mortality rate is very low during the first years of life. Therefore it has potential to radically distort consideration of the surviving hope in the terms of presumption that the population with high infant mortality rate will have necessarily small ratio of older people. Overestimation of life expectancy interpretation relation with the population health condition and mainly with healthcare system functionality represents another disadvantage or frequent mistake at such interpretation. If we didn´t take in account death of healthy people resulted from self-murder, drowning, injuries or other causes, then we would be close to the truth. Demography usually works with mortality rates regardless the death cause.

Picture No. 10 indicates that the life expectancy at birth has been gradually prolonged mainly at men, increasing during monitored period from 69.5 to 72.5 years,



>plot(LE_SR$rok[1:12] , type = „n“, col=“red“, , ylim=c(70,85), xlab = „Roky“, ylab = „Life expectancy“) # empty area
>lines(LE_SR$LE[13:24], type=“o“, pch=21, col=“blue“) # men´s values
>lines(LE_SR$LE[1:12], type=“o“, pch=22, col=“red“) # women´s values
>legend(„bottomright“, c(„Women“,“Men“), cex=0.8, col=c(„blue“,“red“), pch=21:22, lty=1:2, title=“Legend“) # legend

Picture No. 10 Life expectancy in Slovakia in years 2001 - 2012 for both genders. Source: ŠÚSr [20]



Kraj	Roky	
	2001	**2012**
BA	71.48	74.09
TT	69.49	72.45
NR	68.79	71.57
TN	70.59	73.34
ZA	68.89	71.80
BB	67.99	71.32
KE	68.34	71.18
PO	69.68	72.14

Picture No. 11 Differences in life expectancy at men´s birth during 2001 – 2002 according to Slovak regions. Source: ŠÚSr [21]

i.e. by 3 years. At women, reported increase was smaller, from 77.5 years in 2001 to 79.5 years in 2012, i.e. two years of difference. As well, the difference between men and women in the last observed year remained 7 years in the terms of shorter life expectancy for men´s population. Comparison according to particular regions is also very interesting. Values of the life expectancy at birth according to Slovak regions and districts are available in the regional statistical data DB. [21] The following picture shows the comparison in the form of histogram for men and particular Slovak regions. (Picture No. 11).

Despite of the first view that indicates minimum difference, it represents a few years. While the Banska Bystrica region reported the shortest life expectancy in 2001, the Košice region reported such results 11 years later. Difference between Bratislava and Košice region refers to three years, which is a huge difference in men´s average life length between the two regions. We should also take in account the life expectancy increase during 11 years. Of course, the causes of such results can be subject to discussion but such findings deserve our attention.

Summary

This chapter was dedicated to distortion that is eliminated through data standardization or application of life expectancy. We described two methods;

thereof mainly the first one is used at age-based standardization of measures for the compared populations. We presented the method using the library environment {R}. We dealt in detail with the issues of bias and confounding variables that are important for research in various areas of causal relations examination not only in health-related sciences. We briefly mentioned the concept of standard population, as presented by WHO, as well as the characteristics of population dynamics based on mortality rate. Life expectancy is one of frequent characteristics of population when considering and comparing health of various populations. Simple structure and ability of population characterization with the only figure is its major advantage but also restriction. The figure represents the history of life and death of a population that is not here anymore. Thus, considerations that used to be drawn from this indicator should be confronted with further indicators, mainly taken from the studies dealing with health condition of live population. All by all, the life expectancy provides a very useful orientation and the dynamics of the life expectancy development characterizes health behavior of population. Becoming familiar with the mentioned concepts, the reader should be able to understand and interpret the data provided by many organizations.

References

1. Hanley, J. *Edwin Chadwick and the poverty of statistics.* Med Hist, 2002. 46(1): p. 21 – 40.
2. Choi, B. C., de Guia, N. A., Walsh, P. *Look before you leap: stratify before you standardize.* Am J Epidemiol, 1999. 149(12): p. 1087 – 1096.
3. Ahmad, O. et al. *Age standardization of rates: a new WHO standard*. WHO EIP/GPE/EBD, 2001,14.
4. Kahn, H. A., Sempos, C. T. *Statistical methods in epidemiology*. 1989: Oxford University Press.
5. Miettinen, O. S. *Estimation of relative risk from individually matched series.* Biometrics, 1970. 26(1): p. 75 – 86.
6. Last, J. M. *A Dictionary of Epidemiology*. 1995, Oxford: Oxford University Press. 196.
7. Sackett, D. L. *Bias in analytic research.* J Chron Dis, 1979. 32: p. 51 – 63.
8. Delgado-Rodrıguez, M., Llorca, J. Bias. J Epidemiol Community Health, 2004. **58**: p. 635 – 641.
9. Vineis, P. *History of bias*, in *History of epidemiology methods and concepts*, B. Morabia, Birkhauser, Editor. 2004, Springer. p. 405.
10. *Standardization: a classic epidemiological method for the comparison of rates.* Epidemiol Bull/PAHO, 2002. 23(3): p. 9 – 12.
11. Gordis, L. *Epidemiology*. 4th ed. 2008: Saunders.
12. Aragon, T. J. *Epidemiology Tools*. 2010.
13. Anderson, R. N., Rosenberg HM. Age st*andardization of death rates: implementation of the year 2000 standard.* Natl Vital Stat Rep, 1998. 47(3): p. 1 – 16, 20.
14. Chiang, C. L. *The Life Table and its Applications*. 1984, MAlaBAR,FLORIDA: RoberT E. Krieger Publishing Company. 289.
15. Jurčová, D. *Dictionary of demographic terms.* 2005, INFOstat – INSTITUTE OF INFORMATICS AND STATISTICS. Research Center of Demographics, Bratislava.

16. Wynder, E. L. *A corner of history: John Graunt, 1620 – 1674, the father of demography.* Prev Med, 1975. 4(1): p. 85 – 88.
17. Haas, F. E*dmond Halley (1656 – 1742) – Astronomer and Vital Statistician.* JAMA, 1965. 193: p. 66.
18. Mészáros, J. (2000) *Calculation of mortality tables and of lost life years by death, p.* 14.
19. Mészáros, J. (2006) *Structure of of mortality tables*. Forum Statisticum Slovacum, p. 45 – 52.
20. Slovak Statistical Office. *Demographics.*. 2010 [cited 2010 28. 08.]; Available from: http://portal.statis-tics.sk/showdoc.do?docid=2434.
21. Slovak Statistical Office. *Regional Statistical Database. RegDat*. 2013 [cited 2013 22.5.]; Available from: http://px-web.statistics.sk/PXWebSlovak/.

Chapter 7

Study of human population reproduction

Chapter objectives

A reader reasonably asks why it is necessary to deal with demographical statistics if the statistics of health is a primary goal of the publication. The answer results from the demography definition; demography studies human population reproduction and conditions of this process. Demography (in Greek language *demos* – people, *grafein* – describe) is a social science dealing with the study of human population reproduction. [1] Demographical study subject refers to human populations and their reproduction understood as a continuous renewal of human populations; as a consequence of births and deaths. [2] This natural renewal is called a natural exchange or natural movement of population. Demographical events are associated with the demographical reproduction process. Along with birth, death and abortion, demographical events include marriage, divorce, losing a spouse, illness, etc., since these events directly influence the process of natality and mortality. Demography studies these events as mass occurring events. The data taken from the records are at first methodically adjusted to the processes of natality, mortality, marriage, divorce and illness rate and then the processes are analyzed in order to identify their periodicity, short-term variations and long-term trends.

The basis of our interests lays here since the Statistics of Health and Illness studies also the effects that are closely associated with headcount of citizens living in the territory of developed, diagnosed and treated diseases. We are aimed at describing the status and development of group of people being in certain stage of health or health disorder; thus we have to consider the facts related to birth and death of population groups. Here we use the demographical statistics as a tool of acquaintance with these effects. In this chapter, we will deal mainly with the statistics of birth, death and movement of people in various health stages. We will point out to main and derived indicators and methods of their interpretation.

Discovering the surrounding effects, formulation of theories why it is so and verification of the theories on the facts refers to most frequently performed activities of any scientific work. Theory of demography enables such research on the level of demographical processes of birth, death and migration, and their interaction with social, economic, political and cultural processes of recognizing broader relations of a human life and health/ illness correlation. Cognition also enables to understand the consequences of population growth in order to eliminate unwanted effects and influence the future changes. Population growth can get complicated, deepen or even create various social, economic and political problems. Population growth-related problems include safety of food, unemployment, environmental degradation, increased habitation need, need for energies, healthcare and educational facilities and freedom of an individual. Demography can contribute to planning of social, economic, and community services on national and local level.

Historical view

History documents close correlation between the demography and public health. Let´s remind John Graunt (1620 – 1674) who significantly contributed to the health statistics but mainly to demography development. He entered the annals with the analysis of statistics of London citizens and the book he published influenced demographists of that era, and we have been drawing from it to date. Despite of lack of formal education, he was admitted in the Royal Society of England after publishing the book. The society comprised of the top scientists of the era. Graunt lived in England where the *Bills of Mortality* had been published from the start of the 16th century. Death cases were entered in these bills in relation to their residence, religion and disease. It was the era of epidemics when, for example, one fourth of English citizens died in 1625, many of them of pest. The Bills of Mortality were published once per a week and complete annual bill was published at the end of year. Graunt studied these bills and d them with the natality data obtained from the churches. The fact that he recognized major trends in population, for example more boys than girls born and higher mortality rate at men than at women, represents his major merits. He was the first one who described the trends of some diseases development; he estimated London citizen´s headcount, described certain diseases development trends and observed that women attend physicians more frequently than men. Majority of his procedures has been used to date but they are less laborious because of computer use.

Johann Peter Süssmilch (1707 – 1767), Prussian pastor, was another historical personality from the times of demography foundation. He also studied population through statistical methods and tried to demonstrate certain correlation of population effects, considering “God´s Order” its cause. He published the summary of his discoveries in 1741 in the form of book called *Die gottliche Ordnung*. He worked through huge amount of demographical data at the time, searching for common aspects; he recognized the balance between births and deaths and developed the life tables used even in the 19th century.

Great Britain had one more recognized demographist and economist – Thomas Robert Malthus (1766 – 1834). He was known for his pessimistic but very influential opinions related to population growth and its consequences. He supported his conclusion with the idea of uncontrolled growth of human population that will result in lack of food in certain stage. He presented a solution – the morale of an individual and his wise conduct; for example, he recommended marrying in older age and abstaining from sexual contacts. His conclusions were very influential and resulted in changed standpoint to fertility that had been considered before an economic-positive aspect. (enough workforce). His opinions are currently criticized but Malthus remains one of major personalities of the history of demography. [3]

History of the demography started to write deep at the beginnings of recorded human history. According to historian Herotodos, census was conducted in Egypt already around 2900 BC and later during the antic culture in the Antic Rome during the republican era (510 – 529 BC) periodically on five-year basis.

The oldest reliable data about population in the Czech lands are dated in 1754 (in Slovakia it was in 1787). The first census was organized in Austrian-Hungarian territory in 1857. Worldwide census was dated in the beginning of the 70s of the previous century and covered more than four fifth of the world population.

Demography and demographical statistics

We defined demography at the beginning of the chapter as a word based on “population”, however we didn´t explain its meaning. Population (*populus* – people in Latin language) is a societal system that consists of human individuals. They are characterized with individual features, occupy certain space and reproduce. Dictionary of epidemiology [4] specifies general population as all members of human population defined on basis of geographical location, namely country, region, town, etc.

Accordingly, we can put aside particular demographical sections: demographical statics (status and structures of population), demographical dynamics (marriage, divorce, abortion rates, mortality and migration), and demographical prognosis (projections, extrapolations). Demographical statistics is a separate category and we will deal with it later.

Demographical statistics

Demographical statistics deals with processing of quantitative data about the status, structure and movement of population. Subject of research refers to headcount of citizens according to age, gender and other characteristics, as well as marriage, divorce, birth and mortality. It represents the basis of health and disease statistics and serves also for healthcare statistics, since the above stated indicators change with citizens´ headcount in various age groups, and mainly interpret differently. When estimating the needs, it is important to predict the population change dynamics, since the required services also depend on the population, thus also on headcount of newborns, dead, immigrants or emigrants. As we will see later, it is necessary to consider various age structures of population in order to compare health or illness of population in various groups, resulting from demographical statistics.

The expression „demographical“ and „statistics“ is about interconnection of demography and statistics. In the publication, we are not going to describe particular processes of bio-statistics, since we consider that a reader has been acquainted with these fundamentals; otherwise we recommend reading our book about bio-statistics [5] or any of the similar sources that explain fundamental statistical terms.

Newborns

Natural change

Natality statistics represents a part of the statistics of natural changes, describing a part of reproduction cycle. Both events leading to natural change at the population, i.e. birth and death of an individual are recorded in special book known as the Birth Register. All births (live and dead born children) in the specified territory are recorded in the Birth Register, as well as all marriages. Originally, these social events were recorded by the church, in Slovakia it was from the 16^{th} century. In the Habsburgh Monarchy, the state assumed responsibility for keeping the records since the half of the 18^{th} century. Nowadays, the Birth Registers are kept at municipal offices.

Birth report

The Birth Register obtains the information on a newly born child from the processing of Birth Report (*OBYV 2-12)*(Picture No. 1). Healthcare facility where the child was delivered or the child and mother were treated shall fill in the Report in two copies, one for the Birth Register Office and another one as a copy. Before sending the statistical report to The Slovak Statistical Office, the Birth Register shall verify

ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

Registrované ŠÚ SR Č. Vk 6/09
z 10. 6. 2008
Internet: http://www.statistics.sk

Obyv 2 – 12

HLÁSENIE O NARODENÍ

Zdravotnícke zariadenie zašle hlásenie prvý pracovný deň po narodení dieťaťa 2x matričnému úradu.

IKF				Rok		Mesiac	
0	7	0	2				

Hlásenie pre Štatistický úrad SR (doplňte Áno, alebo Nie) */ ....

Ochranu dôverných údajov upravuje zákon č. 540/2001 Z. z. o štátnej štatistike.
Ochranu osobných údajov upravuje zákon č. 428/2002 Z. z. o ochrane osobných údajov.
Za ochranu dôverných a osobných údajov zodpovedá Štatistický úrad SR.

Vyplní matrika:
Matričný úrad ....
Poradové číslo matričné ....

Vyplní zdravotnícke zariadenie:
Okres ....
Miesto narodenia (obec) ....
Spisová značka ....
Pôrodná kniha ....

316. modul | **Narodenie dieťaťa**

Meno a priezvisko dieťaťa */ ....

Poradové číslo (vypĺňa pracovisko ŠÚ SR)	**Dátum narodenia** hodina, minúta ____			**Rodné číslo dieťaťa** */ (vypĺňa matrika)	**Štátne občianstvo**	**Pohlavie** 1 - mužské 2 - ženské
	deň	mesiac	rok			
					uveďte slovom	

Údaje o rodičoch dieťaťa */	Otec	Matka
Meno	....	....
Priezvisko	....	....
Rodné priezvisko	....	....
Dátum narodenia (deň, mes., rok) */	....	....
Miesto narodenia okres	....	....
Miesto narodenia obec	....	....
Štátne občianstvo	....	....
Rodné číslo */		

Dohoda rodičov o priezvisku a mene dieťaťa */

Priezvisko ....

Meno ....

.... podpis otca

.... podpis matky

.... pečiatka zdravotníckeho zariadenia podpis osoby povinnej hlásením

V.......... dňa ....

Záznamy matriky */

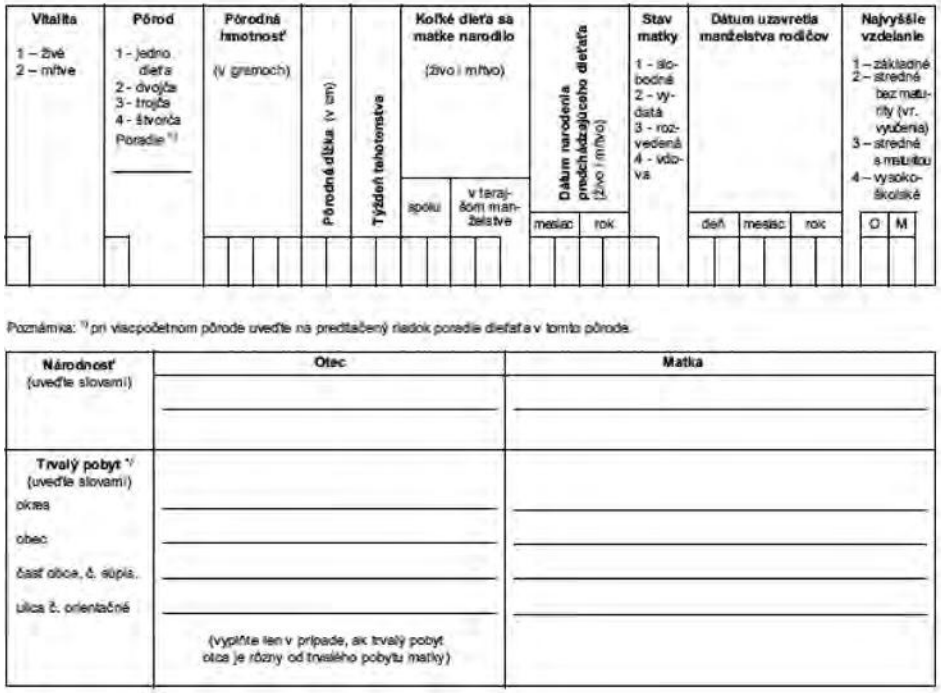
Rodný list vyhotovený dňa .... zaslaný poštou dňa ....

prevzal (dátum, podpis) ....

V .... dňa ....

.... podpis matrikára (-ky)

Pokyny na vypĺňanie sú na druhej strane Hlásenia.



Vitalita 1 – živé 2 – mŕtve	Pôrod 1 - jedno dieťa 2 - dvojča 3 - trojča 4 - štvorča Poradie [1]	Pôrodná hmotnosť (v gramoch)	Pôrodná dĺžka (v cm)	Týždeň tehotenstva	Koľké dieťa sa matke narodilo (živo i mŕtvo)		Dátum narodenia predchádzajúceho dieťaťa (živo i mŕtvo)		Stav matky 1 - slobodná 2 - vydatá 3 - rozvedená 4 - vdova	Dátum uzavretia manželstva rodičov			Najvyššie vzdelanie 1 – základné 2 – stredné bez maturity (vr. vyučenia) 3 – stredné s maturitou 4 – vysokoškolské	
					spolu	v terajšom manželstve	mesiac	rok		deň	mesiac	rok	O	M

Poznámka: [1] pri viacpočetnom pôrode uveďte na predtlačený riadok poradie dieťaťa v tomto pôrode.

	Otec	Matka
Národnosť (uveďte slovami)		
Trvalý pobyt [1] (uveďte slovami) okres obec časť obce, č. súpis. ulica č. orientačné	(vyplňte len v prípade, ak trvalý pobyt otca je rôzny od trvalého pobytu matky)	

Picture No. 1 Birth Report (OBYV 2-12) Source: SOURCE: STATISTICAL OFFICE .

the authenticity and complexity of the data entered by the healthcare facility during the child delivery. Birth report contains information on identification of a newborn (birth date, gender, and nationality), basic characteristics of the delivery (vitality, delivery type, birth weight and length, pregnancy duration in weeks, order of the child, birth date of previous child of the mother), information on the child parents (education, nationality, permanent address and marriage date) and mother´s marital status.

Delivery doesn´t have to be accomplished with a birth of live child. Thus, for statistical purposes, it is necessary to define “live birth” and “dead birth”. Live birth refers to delivered child with at least one sign of life, delivery weight min. 500 g or 499 g or less in case of survived 24 hours after delivery. [6] Breathing, heart response, navelwort pulsation or active muscle movement are considered signs of life at the delivery of child (despite of not cut through navelwort or not expelled placenta). Headcount of live births during certain period represents a sub-variety of total births.

Headcount of newborns includes all children regardless live or dead births. Live and dead births represent its sub-variety. Table No. 1 contains headcount of newborns in Slovakia in 2009 according to status and gender.

Newborns					Live births						Dead births					
in total	CH		D		in total		CH		D		in total		CH		D	
	In total	%	In total	%	in total	% z narode-ných	in total	%	in total	%	in total	% of new-borns	in total	%	in total	%
61445	31702	52	29743	48	61217	100	31563	52	29654	48	228	0	139	61	89	39

Table No. 1 Headcount of newborns, live and dead births in Slovakia in 2009. Source: Slovak Statistical Office: Stats and movement of citizens in Slovakia in 2009. http://portal.statistics.sk/showdoc.do?docid=17312

Headcount of born boys is by 4% higher than girls. This will change during the first years of life. Fortunately, there aren´t many dead-born girls - less than 1% of total newborns (0.37%, i.e. less than 4 pro mil). There are more dead born boys than girls.

Along with absolute figures (number of events), relative or dependant events are used, i.e. dependant on other statistical event. Birth, *natality*, or *fertility* describes children delivery as a mass demographical effect. [6] Natality completed with *fertility* represents accomplished physiological fertility and it plays major role in reproduction process of any population.

Natality is very ambiguous term. A few better defined terms have been used in the demography; thereof the following are closest to the commonly understood term of natality: Gross Natality Rate, Natality, Nativity – simple ratio of live births and the size of population. It is calculated per a population median in particular year, i.e. population headcount as of July 01 in particular year; and usually expressed in pro mil, i.e. per 1000 citizens and marked as *gross live birth rate.* If all newborns are taken in account (live and dead births), it is marked as *gross total natality rate.* [6] *Net reproduction rate* is an expression of population growth. If equal to 1, the population size doesn´t change. If smaller than 1, the population size decreases, or increases in case of net reproduction rate higher than 1. *Aggregate fertility* represents headcount of children that is delivered by one woman in average. Among the natality related indicators, there are two interesting ones: *average mother age at delivery and average mother age at the first delivery.*

Dynamics of birth

Gross natality rate (Table No. 2) and other measures are mainly used for research of population dynamics and are rather useless when making research in the area of public health and epidemiology. Headcount of newborns is usually a significant parameter in estimation of reproduction health program efficiency and for

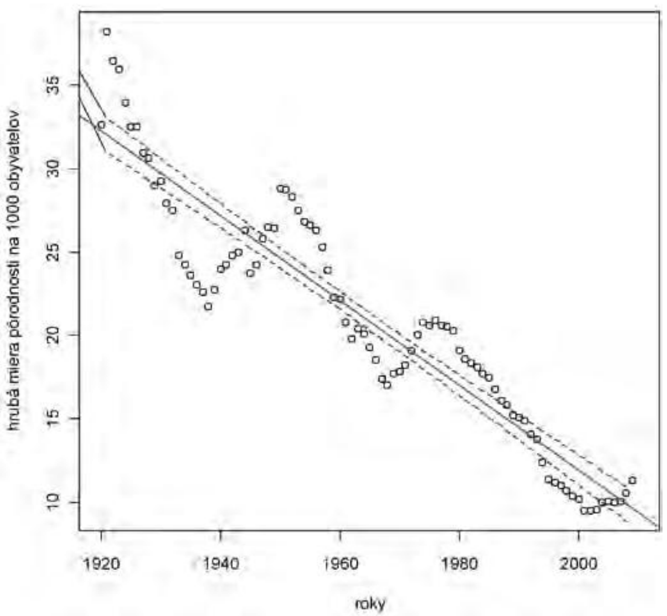
Preliminary median (July 01, 2009)			Live births			Gross natality rate		
in total	men	women	in total	men	women	in total (‰)	men (‰)	women (‰)
5416 958	2632702	2784256	61217	31563	29654	11	12	11

Table No. 2 Gross natality rate. Source: Slovak Statistical Office: Stats and movement of citizens in Slovakia in 2009.. http://portal.statistics.sk/

> summary(regres_porod)
Call:
lm(formula = hm_porod ~ roky)
Residuals:
Min 1Q Median 3Q Max
-- - -- -
5.9738 1.7566 0.0561 1.7657 6.2170
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	518.92242	20.06780	25.86	<2e-16	***
roky	-0.25348	0.01021	-24.82	<2e-16	***

---
*Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1*
Residual standard error: 2.517 on 88 degrees of freedom
Multiple R-squared: 0.875, Adjusted R-squared: 0.8736
F-statistic: 615.9 on 1 and 88 DF, p-value: < 2.2e-16



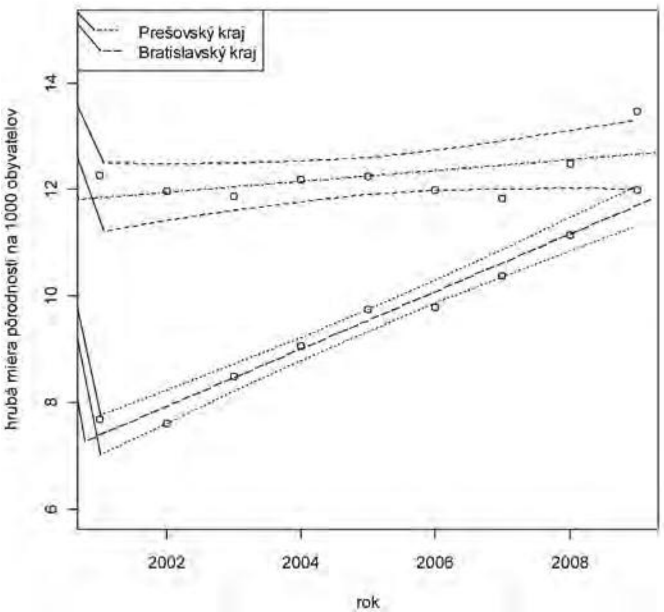
Picture No. 2 Development of gross natality rate in Slovakia(1920 – 2009). Source: SOURCE: STATISTICAL OFFICE, 2011 specification of service needs. In this case it is mostly about services provided to a woman during pregnancy, for example preventive examinations or pre-delivery hospitalization, assistance during delivery as well as estimation of health and healthcare service cost associated with woman and child. Data applicable to the Slovak Republic can be drawn directly from the Slovak Statistical Office website and organization founded by the Office, - INFOstat[23]. Any statistical value applicable to a moment in a time (year, month, day) doesn´t provide us with many reasons for research. Considerations of the situation in a community, territorial whole or country are made mainly on basis of identified development of headcounts or identificator in a time and in comparison to the other ones (neighbors, larger territorial whole or the world). Naturally, such procedures use statistical methods that will be gradually described on the examples. At first, we will demonstrate development of natality rate in Slovakia during period of years 1920 - 2009 (Picture No. 2) and compare it to the development in two districts of Slovakia (Picture No. 3). Development of gross natality rate was presented in particular years. Then we plotted the regression line and the interval of certainty corresponding to 95 %. Regression line matches actual values as proved by value R^2 that approximates 1.

Bratislava region

> summary(regr_BA)
lm(formula = BA ~ rok)
Residuals: Min 1Q Median 3Q Max
-- - 0.3247 - - 0.2413 0.0280 0.2033 0.3027
Coefficients:
Estimate Std. Error t value Pr(>|t|)
-- - 9.58e-- -
*(Intercept) 1.068e+03 6.730e+01 -- - 15.87 07****
5.373e-- - 3.356e-- - 9.01e-- -
*rok 01 02 16.01 07****
Residual standard error: 0.26 on 7 degrees of freedom
Multiple R-- - squared: 0.9734, Adjusted R-squared:
0.9696
F-statistic: 256.3 on 1 and 7 DF, p-value: 9.013e-07

Prešov region

Adjusted R-- -
>summary(regr_PO)
Call:
lm(formula = PO ~ rok)
Residuals: Min 1Q Median 3Q Max
- 0.61989 -- - 0.18122 -- - 0.00556 0.03661
0.80578
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -- - 192.58861
115.25918 -- - 1.671 0.139
year 0.10217 0.05749 1.777 0.119
Residual standard error: 0.4453 on 7 degrees of freedom
Multiple R-- - squared: 0.3109, Adjusted R-- - squared: 0.2125
F-- - statistic: 3.159 on 1 and 7 DF, p-- - value: 0.1188
Multiple R-- - squared: 0.9734, Adjusted R-- - squared: 0.9696
F-- - statistic: 256.3 on 1 and 7



Picture No. 3 Development of gross natality rate in Bratislava and Prešov region during 2000 – 2009.

Source: Statistical Office SR:

In the following tables (Table No. 3 and 4) you can find the process in {R} environment including the comments. Calls in any library are not necessary to draw up the graphs. Line color can be changed for better visualization.

 # 1st step – preparation of variables where we enter the years in from the 1st column of
 # Excel table and 2nd column therefrom in the variable gross_natality_rate
 # We work in the environment RExcel.
 # View on ready data is reduced since the columns are long. We take the fields with years in a block and click on the right mouse button, choosing „Put R Var“; then we enter either the variable name in the field called „Array name in R“ or enter it, clicking on the applicable field. In {R} environment, we can check whether our variable was initiated.
 # Plotting gross natality rate development points: bodov
> plot(rok,hruba_miera_porodnosti,xlim=c(1920,2010),main=(„Vývoj hrubej miery pôrodnosti v SR“),ylab=(„hrubá miera pôrodnosti na 1000 obyvatelov“))

 # calculation and plotting of linear regression line
> abline(lm(hruba_miera_porodnosti~rok), col=“black“)

```
# we enter sequence of years 1920 – 2009 in the variable
newx > newx<-seq(1920,2009)

# we enter prediction result with the certainty interval in the variable pred
> pred<-predict(lm(hruba_miera_porodnosti~rok),newdata=data.frame(x=newx),interval =
c(„confidence“), level = 0.95,type=“response“)

# we plot the certainty interval lines
> lines(newx,pred[,2],col=“blue“,lty=2) # horný interval
> lines(newx,pred[,3],col=“blue“,lty=2) # dolný interval
```

Table No. 3 Process of gross natality rate in Slovakia plotting in the graph. Source: Statistical Office SR, 2011

```
# in RExcel environment, we gradually  choose variable from the table in the column year, PO,
  BA in the way similar to the previous table  ( Error! Reference source not found.).
# we plot results and values for Bratislava region
> plot(rok,BA,main=(„Gross natality rate development in selected regions“), ylab=(„
Gross natality rate per 1000 citizens“),ylim=c(6,15))
> par(new=T)

# in the same way, we plot values for Prešov region
>plot(rok,PO,ylab=(„ Gross natality rate per 1000 citizens“),ylim=c(6,15))

# then we calculate regression for Prešov and Bratislava
region against years >
abline(lm(PO~rok),col=“black“,lty=4)
> abline(lm(BA~rok),col=“black“,lty=5)

# we enter years in a new variable and call the function predict() determining certainty intervals on
  95 level, plot them at first for Prešov and then for Bratislava region
> newx<-seq(2001,2009)
>  pred<-predict(lm(PO~rok),newdata=data.frame(x=newx),interval                    =
c(„confidence“),level = 0.95,type=“response“)
> lines(newx,pred[,2],col=“black“,lty=2)
> lines(newx,pred[,3],col=“black“,lty=2)
> pred_BA<-predict(lm(BA~rok),newdata=data.frame(x=newx),interval                  =
c(„confidence“),level = 0.95,type=“response“)
> lines(newx,pred_BA[,2],col=“black“,lty=3)
> lines(newx,pred_BA[,3],col=“black“,lty=3)
> legend(„topleft“, c(„Prešovský kraj“,“Bratislavský kraj“),lty=c(4,5))
```

Table No. 4 Plotting gross natality rate in the graph for Prešov and Bratislava region. SOURCE: STATISTICAL OFFICE , 2011

Birth comparison

We can compare counts, indicators or statistical parameters. Any review made on basis of such comparison bears bias resulting from possible differences in compared populations, e.g. the age structure. Thus it is necessary to ensure elimination of such source of inaccuracies with standardization. Such procedure is described in a separate chapter.

Abortion statistics

Abortion is a premature self-induced or induced termination of pregnancy when the fetus doesn´t show vital signs and its birth weight is lower than 1000 grams, or if fetus shows some vital signs but its weight is below 500 grams and it doesn´t survive 24 hours, or the fetus weight cannot be measured and the pregnancy took less than 28 weeks. Regulation No. 22/1988 Coll. defines the term in Slovakia or other definitions in other documents. There is no international definition of abortion. [6] Abortion data are compiled by the National Center of Health Information, before it was done by the Health Information and Statistics Institute in Bratislava in the form of processing the *Application for Artificial Pregnancy Termination and Abortion Report.* The state statistics takes it from the IS kept by the Slovak Ministry of Healthcare. [7]

We distinguish between spontaneous, self-induced and artificial / induced abortion that can be named also interruption. Spontaneous abortion (fetal death, miscarriage) is a biologically-induced abortion without apparent external intervention. On the contrary, induced abortion refers to artificially induced pregnancy termination. The statistics usually records legally performed abortions.[24]

abortions	**Self-induced abortions**	**UPT[24] up To 8th week**	**UPT 9.–12. w.**	**UPT 13.-24.w.**	**exfoetation**	**other UPT type**	**illegal UPT**
17935	4695	6706	3071	193	382	2888	0

Table No. 5 Abortion statistics in Slovakia in 2009

Preliminary median (1. 7. 2009)	**Abortions**			**Gross abortion rate**		
in total	in total	Self-induced	UPt	in total	Self-induced [‰]	UPt [‰]
5416958	17935	4695	12858	3	1	2

Table No. 6 Gross abortion rate in Slovakia in 2009

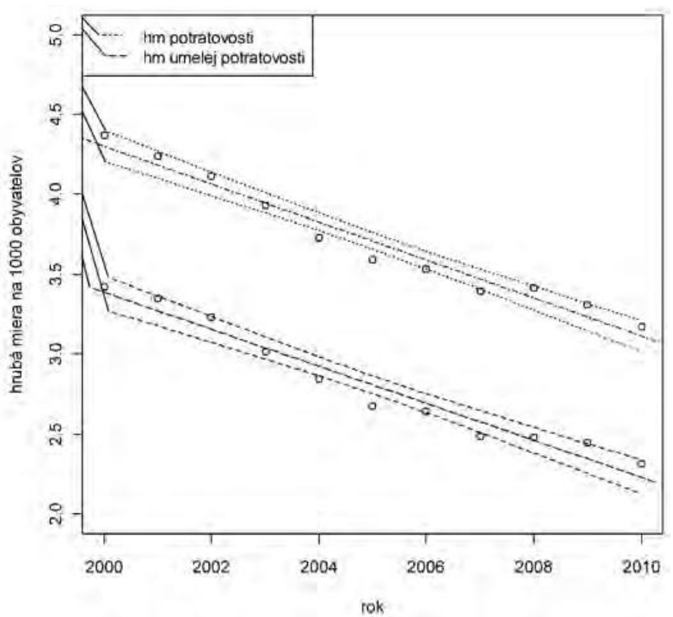
24 UPT – induced abortion

Gross abortion rate represents the number of abortions vs average population per a year. It is usually expressed in pro mil. and stated separately per self-induced and induced abortions.

General abortion rate allocates number of abortions to the average headcount of women in fertile age (15 - 49 years of age), usually per a year. It is usually expressed in pro mil. and stated separately per self-induced and induced abortions. We will mention the abortion ratio as another indicator expressing the number of abortions per total count of newborns per a year. It is usually expressed in per cent and stated separately per self-induced and induced abortions. It sometimes includes also live births.

> abortion
Call:
lm(formula = hm_potrat ~ rok)
Residuals:
Min 1Q Median 3Q Max
-0.12011 -0.06618 0.05095 0.05869 0.07726
Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 242.192917 15.027874 16.12 6.04e-08 ****
*rok - 0.118945 0.007495 -15.87 6.91e-08 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: 0.07861 on 9 degrees of freedom
Multiple R-squared: 0.9655, Adjusted R-squared: 0.9617
F-statistic: 251.8 on 1 and 9 DF, p-value: 6.909e-08

> induced abortion
Call:
lm(formula = hm_umely ~ rok)
Residuals:
Min 1Q Median 3Q Max
-0.13164 -0.06675 0.01300 0.07852 0.09451
Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 233.673961 16.273844 14.36 1.65e-07 ****
*rok -0.115144 0.008117 -14.19 1.83e-07 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: 0.08513 on 9 degrees of freedom
Multiple R-squared: 0.9572
Adjusted R-squared: 0.9524
F-statistic: 201.2 on 1 and 9 DF, p-value: 1.829e-07



Picture No. 4 Gross abortion rate in Slovakia. Source: Statistical Office SR

Age specific abortion rate compares total abortions of women in particular age and the average women´s headcount of the same age. It is usually expressed in pro mil. and stated separately per self-induced and induced abortions. Sometimes it is called *specific abortion rate*.

Gross abortion rate development graph illustrates reduction of both rates during the last decade (Picture No. 4). The trend of re-plotted regression lines is favorable ($R^2 = 0.96$ and 0.95). The abortion-related data disallow for identification of the trend causes. We can suppose that the trend corresponds to the political and economic development in Slovakia, and to the changed abortion pills availability, or to health education. The following table contains statistical data calculation and plotting method. (Table No. 7).

```
# We chose the numbers of aborts and induced aborts for period 2000 – 2010 from the data available on
  SLOVSTAT On Line
# we entered them in EXCEL file and created the data frame, compiling them in a block and
choosing „Put R Data Frame“. The structure was named „abortions“. Using the function
attach() we made particular variables accessible. Further we proceeded analogically to the
previous cases.
> attach(potraty) # extraction of variables from data frame "abortions"
# plotting both variables
> plot(rok,hm_potrat,main=(„Gross abortion rate development in Slovakia“),ylab=(„gross
rate per 1000 citizens“),ylim=c(2,5))
> par(new=T)
> plot(rok,hm_umely,,ylab=(„ gross rate per 1000 citizens “),ylim=c(2,5))
# relocation of lines through both variables values
> abline(lm(hm_potrat~rok),col="black“,lty=4)
> abline(lm(hm_umely~rok),col="black“,lty=5)
# calculation and plotting of certainty intervals
> newx<-seq(2000,2010)
> pred<-predict(lm(hm_umely~rok),newdata=data.frame(x=newx),interval =
c(„confidence“),level = 0.95,type="response“)
> lines(newx,pred[,2],col="black“,lty=2)
> lines(newx,pred[,3],col="black“,lty=2)
> pred_potrat<-predict(lm(hm_potrat~rok),newdata=data.frame(x=newx),interval = >
c(„confidence“),level = 0.95,type="response“)
> lines(newx,pred_potrat[,2],col="black“,lty=3)
> lines(newx,pred_potrat[,3],col="black“,lty=3)
# lenend insertion in the picture
> legend(„topleft“, c(„hm potratovosti“,"hm umelej potratovosti“),lty=c(4,5))
# regression parameter per each variable
> potrat <- summary(lm(hm_potrat~rok))
> umely <- summary(lm(hm_umely~rok))
```

year	abortions	Induced abortions	median	Gross rate	
				abortion	Induced abortion
2000	23593	18468	5400679	4.37	3.42
2001	22792	18026	5379780	4.24	3.35
2002	22141	17382	5378809	4.12	3.23
2003	21159	16222	5378950	3.93	3.02
2004	20075	15307	5382574	3.73	2.84
2005	19332	14427	5387285	3.59	2.68
2006	19054	14243	5391184	3.53	2.64
2007	18318	13424	5397766	3.39	2.49
2008	18452	13394	5406972	3.41	2.48
2009	17935	13240	5418374	3.31	2.44

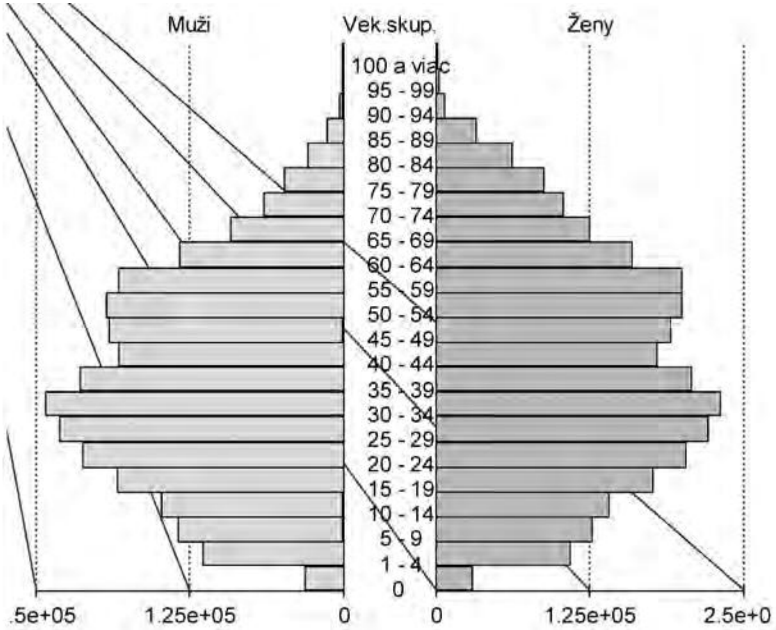
Table No. 7 Process of plotting of gross abortion rate and gross induced abortion rate development in Slovakia

Use of abortion statistics

Monitoring of population headcount has a few goals; one of them refers to its application to identification of health or other disorders leading to self-induced abortions/ miscarriage. [8 – 11] Example of the study about miscarriage and other delivery-related problems is the study conducted in Amsterdam, dealing with the impact of social factors in particular city districts, e.g. income level, unemployment or other social security, on the pregnancy outcome. [12] Along with social factors, effects of physiological [13] and other factors were studied. [14 – 16] Special attention has been paid to the abortions of adolescents. [17] Abortion data are used also in the study about reproduction health program efficiency or impact of the changes resulting from amended legislation. [18 – 20] The study in USA monitors the differences in particular US state legislations and their impact on the abortion rate. [21] National Health Information Center deals with the abortion statistics in Slovakia, issuing a publication on annual basis that deals with the abortion development only. [22] The publication describes current status and development without seeking applicable correlations.

Population headcount and increase

Population as of certain moment is probably historically the oldest and one of the basic characteristics monitored by the demographical statistics. It captures all citizens with permanent address in particular town/ location. It results from the census that includes counting of houses and apartments as well, identifying population headcount and structure according to age, marital status and nationality, which shall serve



Picture No. 5 Age pyramid, Slovakia. Slovak population age structure according to gender and 5-year age groups as of July 01, 2010. Source: Statistical Office SR

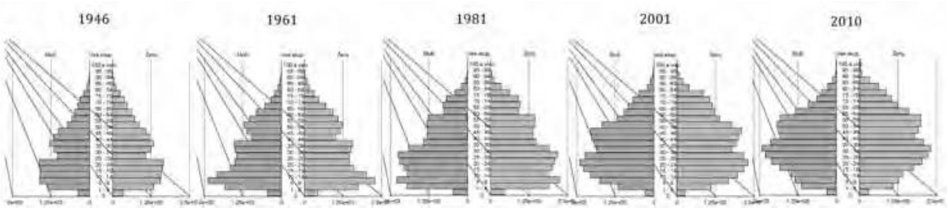
for estimation of population development in the following decade (till the next census). In general, we can say that population and its structure identified through census represents the grounds for health discrepancies and other population health characteristics study. [23, 24]

Population headcount differs in various periods of time. *Initial headcount in a year* represents headcount of particular territory citizens at the beginning of the monitoring period / usually calendar year/. Analogically, *end headcount in a year* is stated (as of Dec ~~31.) that is usually eq~~ual to the *initial headcount in the following year. Mean headcount in a year* [25] is usually reported as of July 01. [6] Population headcount median is used only for calculation of demographical and other statistics (natality or abortion rate – see above). Population structure is distinguished according to age, gender, permanent address or marital status.

[25] Headcount of citizens having permanent address in Slovakia as of June 30 (July 01), in the demographical statistics had been used till 2010 inclusive for calculation of the population median. From 2011, it´s been calculated as an arithmetical mean of headcounts.

Population of particular territory also changes with time. Demographical events that occur in particular year are determining for *population movement*. These events include births, deaths, abortions and movement. Resulting movement is expressed as total increase, calculated as the sum of natural increase and migration balance. *Natural population increase* represents the difference between the number of newborns and deads. *Migration balance* represents the difference between headcount of citizens moved in and out. Negative value of total increase is marked as *total decrease.* [6]

Headcount of live citizens in particular territory can be illustrated also according to age and gender, resulting in so called age pyramid / population pyramid (Picture No. 5). In the example, we used 5-year age groups. Such illustration helps understand the citizens age structure and estimate the population development. Population increase is expected when many children are born and vice versa, with stagnating population somewhere in between. Let´s look at the Slovak population age pyramid development since the WWII. (Picture No. 6). The war consequences can be seen as a notch in the age group 25 – 29 and slightly less at the age group 30 – 34 (both men and women). Such discontinuity was caused by deaths between WWI and WWII that had gradually shifted to higher age groups in the following decades, and finally disappeared. You can also notice the shape in particular decades. It is of a triangle shape in 1946 and 1961with a wide base that sharply narrows towards the top. Such population is characterized with strong natality but also strong death rate from age group 50+.



Picture No. 6 Comparison of Slovak population development within 1946 - 2010.
Source: Statistical Office SRSR

Such shape represents expanding population.

In the 80s, the pyramid base is still wide, referring to acceptable population increase but surviving to older age prolongs slightly and the population is ageing. In 2001, we can notice strong base narrowing, indicating reduced count of newborns, which was even more remarkable in the following decade. Population ageing is even more remarkable but total life expectancy isn´t dramatically prolonged.

DB Slovstat is accessible from website of SR Statistical Office after login. The DB provides for access to statistical data of Slovakia (also historical). Choose “search” from the offer and to find a table, state word “citizen”.
In the table offer containing the word, choose table “Age structure of Slovak population according to gender and 5-year age groups (1945 - 2010)“. The display offers population status option as of particular date (half-year, end year), gender (using SHIFT[26] we chose both genders), and choose all age groups as well as year.
Click on „create table“ and data export in Excel will be allowed.
If you use Rexcel, further process is very easy: At first, install package PYRAMID (details are either in R project manual or in the book „Bio- Statistics for Public Healthcare Students“[5]).

> library (pyramid)

Author of the function provided very detail manual to follow when formulating the function call. At first we have to prepare the data. The function requires input data in three columns; the 1st and the 2nd column with population headcount according to gender and the 3rd one serves for description of age groups.
Thus, the table obtained from SR Statistical Office should be reformatted in required shape through columns copying. Take care since the data are in text form and should be converted to digits.
Fortunately, Excel has such function. Then choose the table in the block and press the right mouse button. In case of properly installed Excel, the offer will pop up, PutRDa-taFrame is selected and the name shall be allocated to the new data structure, e.g. citizen_SR_1951. Then the pyramid function call order is written in any of the empty fields >
pyramid(obyvatel_SR_1951,Llab=“Muži“,Rlab=“Ženy“,
Clab=“Vek.skup.“,Laxis= seq(0,250000,len=3))
# Parameters Llab, Rlab and Clab specify text to appear in the pyramid fields heading and the last parameter L-axis determines data extent and count of dashes on x-axis.
# If all was done correctly, picture will be drawn with age pyramid. It can be saved for later use.

Table No. 8 Process of population pyramid drawing in R environment

[26] RExcel is a program enabling assignment of orders for program R directly from Excel. It supports easy data transfer between two environments.. RExcel is freely downloadable from http://rcom. univie.ac.at/ .

Population pyramid shape is significantly different in the developing countries. It is characterized with a wide base as a result of high natality rate, and sharp narrowing from the middle age that is caused by high mortality rate. Countries with developed economy have usually the pyramid with narrow base and mortality rate affects the pyramid shape at first in older age. Usefulness of such illustration is relative; a public health expert encounters it in the review of health and healthcare services requirements as a result of population ageing or in developing countries where it confirms high mortality rate and high natality rate.

Deaths

Statistics of deaths is prepared in the form of "*Deceased Obduction Form and Death Statistical Report (OBYV 3-12)* (Picture No. 7). Registers of Birth/Death represent the reporting units. Prior to sending the statistical report to the Slovak Statistical Office, the Register shall verify credibility and completeness of the data stated by the physician during obduction/ autopsy of the deceased. The following statistical data are sought for through the Death Report: death date, personal data of the deceased (birth date, permanent address, gender, marital status, nationality and citizenship), death cause, autopsy mark, set of data required at deceased children up to 1 year of age (life duration, birth weight, legitimacy, death site). [7] Headcount of the dead structured according to age and gender represents the basic data from which indicators are derived in the form of mortality rate, and further specified according to age and special population groups.

Gross mortality rate

Gross mortality rate represents headcount of the dead compared to the mean headcount of citizens per a year. It is usually expressed per 100 thousand citizens. This rate is strongly influenced by age structure since the age represents major mortality rate determinant. To make comparison between populations, it is necessary to perform standardization to reference population. If we mark gross mortality rate as CDR_t, where parameter t refers to particular year, and mark the count of the dead in the year as D_t and use N_t as the population mean in particular year t, then the rate is expressed as follows:

$$CDR_t = \frac{D_t}{N_t} \times 100000.$$

ohreb spopolnením sa povoľuje - sa nepovoľuje (nehodiace sa prečiarknite)

....

dňa podpis a pečiatka príslušného orgánu

evoz na pochovanie sa povoľuje - sa nepovoľuje (nehodiace sa prečiarknite)

....

dňa podpis a pečiatka prehliadajúceho lekára

....

anovisko regionálneho úradu verejného zdravotníctva (v prípadoch podľa 3 ods. 3 zákona č. 470/2005 Z. z. o pohrebníctve) dňa podpis a pečiatka

Stav	Rodné číslo pozostalého manžela (-ky)	U detí zomrelých do 1 roka uveďte			
		Pri úmrtí do 24 hodín dĺžka života v hodinách. U starších detí vpíšte „99"	**Kde nastalo úmrtie?** 1 - v ústave 2 - doma 3 - inde	**Pôrodná hmotnosť** (v gramoch)	**Bol otec dieťaťa manželom matky?** 1 - áno 2 - nie
1 - slobodný (-á) 2 - ženatý, vydatá 3 - rozvedený (-á) 4 - ovdovený (-á)	(ak nie je možné zistiť, treba vykonštruovať prvých šesť miest z dátumu narodenia)				

	Príčina smrti podľa klinického nálezu a 4-miestna značka (kód) [5]	**Kód**
I.	a) choroba (stav), ktorá (-ý) priamo privodila (-il) smrť [6]	
	b) predchádzajúce príčiny	
	c) prvotná príčina	
II.	iné závažné chorobné stavy a zmeny	
III.	išlo o pracovný úraz, náhodný úraz, vraždu, samovraždu? .... Podčiarknite a uveďte mechanizmus smrti!	

....

miesto prehliadky deň, mesiac, rok a hodina prehliadky podpis a pečiatka prehliadajúceho lekára

ávrh prehliadajúceho lekára ....

ávrh na pitvu, druh pitvy, zdravotno-bezpečnostné opatrenia, lehota a spôsob pohrebu) podpis a pečiatka ošetrujúceho lekára

	Záverečná diagnóza podľa vykonanej pitvy a 4-miestna značka (kód) [5]	**Kód**
I.	a) choroba (stav), ktorá (-ý) priamo privodila (-il) smrť [6]	
	b) predchádzajúce príčiny	
	c) prvotná príčina	
II.	iné závažné chorobné stavy a zmeny	
III.	išlo o pracovný úraz, náhodný úraz, vraždu, samovraždu? .... Podčiarknite a uveďte mechanizmus smrti!	
ola vykonaná pitva?	Pre zápis do rubriky použite kód: 1 - áno, 2 - nie (vypĺňa pracovisko ŠÚ SR)	

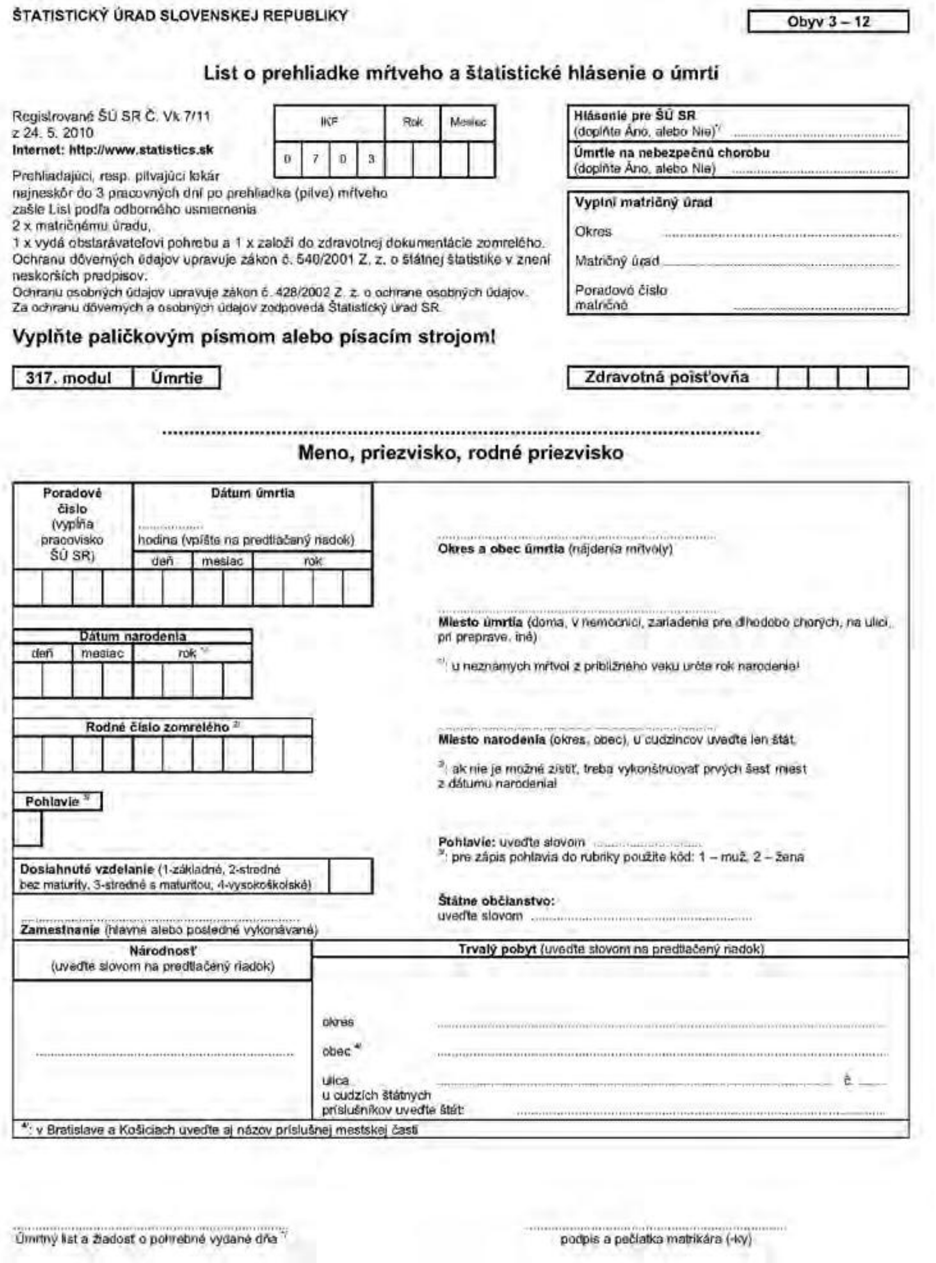
....

miesto pitvy deň, mesiac, rok a hodina pitvy podpis a pečiatka lekára, ktorý vykonal pitvu

Kódy pre štatistické spracovanie ríčin smrti (kóduje pracovisko ŠÚ SR)		***/: Pokyny na vypĺňanie „Listu o prehliadke mŕtveho a štatistického hlásenia o úmrtí" pre ŠÚ SR**
I. až XIX. kap.	XX. kap.	1) List pre ŠÚ SR označte v záhlaví na prvej strane vpravo hore doplnením „Áno". 2) Prvú stranu Listu vyplňte ako prvú kópiu, druhú stranu ako originál. 3) V Liste nevypĺňte údaje matričného úradu „Úmrtný list a žiadosť o pohrebné vydané dňa....".

Medzinárodná štatistická klasifikácia chorôb a pridružených zdravotných problémov v znení 10. decenálnej revízie (MKCH-10)

To neznamená spôsob smrti (napr. zlyhanie srdca), ale chorobu, úraz, komplikáciu, ktorá spôsobila smrť!


ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

Obyv 3 – 12

List o prehliadke mŕtveho a štatistické hlásenie o úmrtí

Registrované ŠÚ SR Č. Vk 7/11
z 24. 5. 2010
Internet: http://www.statistics.sk

IKF				Rok		Mesiac	
0	7	0	3				

Hlásenie pre ŠÚ SR (doplňte Áno, alebo Nie)[1] ....

Úmrtie na nebezpečnú chorobu (doplňte Áno, alebo Nie) ....

Prehliadajúci, resp. pitvajúci lekár najneskôr do 3 pracovných dní po prehliadke (pitve) mŕtveho zašle List podľa odborného usmernenia
2 x matričnému úradu,
1 x vydá obstarávateľovi pohrebu a 1 x založí do zdravotnej dokumentácie zomrelého.
Ochranu dôverných údajov upravuje zákon č. 540/2001 Z. z. o štátnej štatistike v znení neskorších predpisov.
Ochranu osobných údajov upravuje zákon č. 428/2002 Z. z. o ochrane osobných údajov.
Za ochranu dôverných a osobných údajov zodpovedá Štatistický úrad SR.

Vyplní matričný úrad
Okres ....
Matričný úrad ....
Poradové číslo matričné ....

Vyplňte paličkovým písmom alebo písacím strojom!

317. modul	Úmrtie

Zdravotná poisťovňa				

....
Meno, priezvisko, rodné priezvisko

Poradové číslo (vypĺňa pracovisko ŠÚ SR)	Dátum úmrtia .... hodina (vpíšte na predtlačený riadok)		
	deň	mesiac	rok

.... **Okres a obec úmrtia** (nájdenia mŕtvoly)

.... **Miesto úmrtia** (doma, v nemocnici, zariadenie pre dlhodobo chorých, na ulici, pri preprave, iné)

Dátum narodenia		
deň	mesiac	rok[1]

[1]: u neznámych mŕtvol z približného veku určte rok narodenia!

Rodné číslo zomrelého[2]

.... **Miesto narodenia** (okres, obec), u cudzincov uveďte len štát.

[2]: ak nie je možné zistiť, treba vykonštruovať prvých šesť miest z dátumu narodenia!

Pohlavie[3]

Pohlavie: uveďte slovom ....
[3]: pre zápis pohlavia do rubriky použite kód: 1 – muž, 2 – žena

Dosiahnuté vzdelanie (1-základné, 2-stredné bez maturity, 3-stredné s maturitou, 4-vysokoškolské)

Štátne občianstvo:
uveďte slovom ....

....
Zamestnanie (hlavné alebo posledné vykonávané)

Národnosť (uveďte slovom na predtlačený riadok)	**Trvalý pobyt** (uveďte slovom na predtlačený riadok)
....	okres .... obec[4] .... ulica .... č. ..... u cudzích štátnych príslušníkov uveďte štát ....

[4]: v Bratislave a Košiciach uveďte aj názov príslušnej mestskej časti

.... Úmrtný list a žiadosť o pohrebné vydané dňa[1]

.... podpis a pečiatka matrikára (-ky)

Picture No. 7 Obduction Report of Deceased (Population 3-12)

Infant mortality rate

Infant mortality rate represents mortality of children up to one year of age after the delivery. It is a proportion of deaths of newborns up to one year of age to total count of newborns in particular year, recalculated per 1000 newborns. If we mark infant mortality rate as IMR_t, number of deaths of newborns up to one year of age (between year 0 and 1) in the year as D_t and number of all live births B_t in the year t, the proportion shall be expressed as:

$$IMR_t = \frac{D_t}{B_t} \times 1000.$$

Infant mortality rate is often used when characterizing regional or national healthcare. The smaller is the value, the better is the rate. Poor countries have significantly higher rate than rich countries. The rate can vary also within a country, region or community because of various combinations of causes, for example problems during delivery, poor health condition of mother, congenital defects or nutrition. If mother is not able to breastfeed her child, there are not many kids that survive the conditions of insufficient hygiene, lack of clean water and enough nutrition. Improvement of communal hygiene and habitation in Europe caused that the rate has been permanently decreasing since the 2nd half of the 19Th century; including much better care of mother and child by the healthcare system. Therefore the rate has been considered a useful indicator of a country condition and health as well as administration quality. [25]

Average infant rate in EU refers to 5.2 dead children per 1000 live births. [25] The lowest value 2.1 was reported by Comunidad Foral de Navarra (Spain) and the highest value exceeding 20.1 by northern Romania. To compare, the table shows the values of selected countries with visible current remarkable differences. (Table No. 9). It is necessary to use the indicator very carefully and draw conclusions from other indicators as well.

Age-specific mortality rate

Age-specific mortality rate represents a general method of measuring mortality deviation in the terms of age. The rate, marked as M_a, is usually defined as proportion of dead person D of particular age a during particular year, let´s mark them D_a and average citizens´ headcount in particular year, age group a marked as N_a, and usually recalculated per 1000 citizens in particular age group:

$$M_a = \frac{D_a}{N_a} \times 1000.$$

Country	**IMR**
Sub-Sahara Africa	89
Less developed countries	82
World	47
Eastern Europe	11
Europe	7
Slovakia	7
North America	6
USA	6
UK	5
Austria	4
Denmark	4
Spain	4
Switzerland	4
Norway	3

Picture drawing:
> par(mar=c(5,12,4,2))
> par(las=2)

> barplot(dojc, horiz=TRUE, names.arg =krajiny) #dojc – variable with data, countries

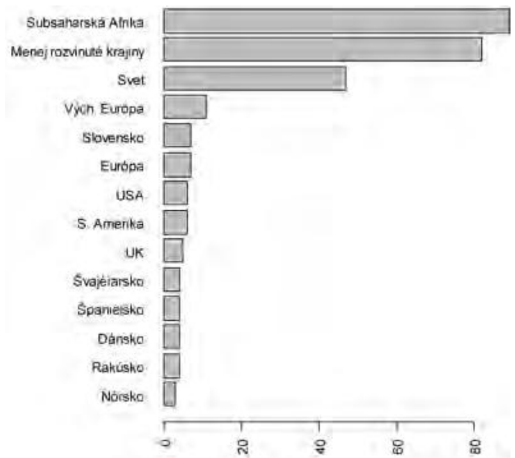
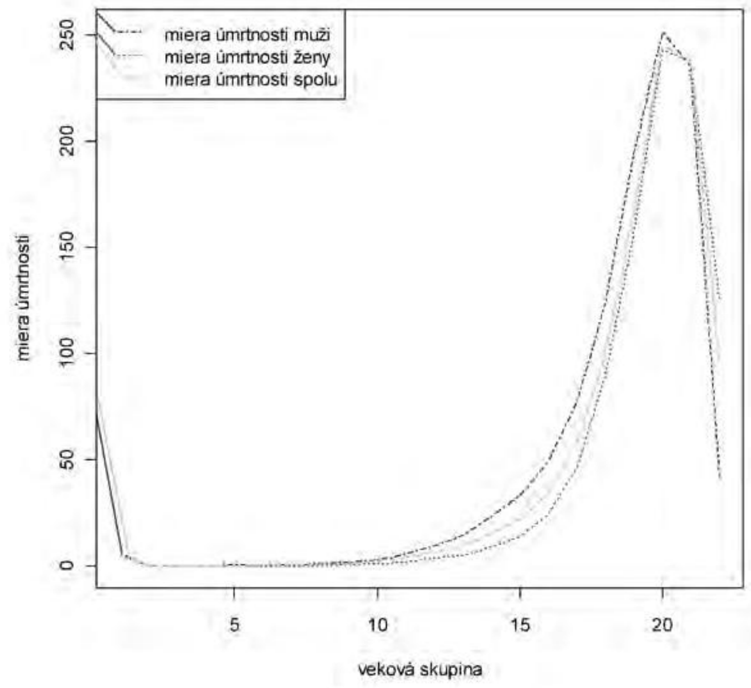


Table No. 9 Infant mortality rate in selected countries during 2005 – 2010. Source: UNDATA

Graphical illustration of the infant mortality rate for all age groups is usually of asymmetrical shape U. Its value is high after birth but decreases to minimum around the age 10 - 14 years. Then it gradually, less or more evenly, increases to maximum. The oldest age groups within 60 - 89 years show the highest mortality rate. The rate is calculated usually per 5 or 10-year age groups but there are two age groups for early age – up to one year and within 1 – 4 years of age because of relatively high infant mortality rate. [26] The rate is often calculated separately per men and women.

Let´s take two populations with the same gross mortality rate but much more kids dying in the first age groups in one than another group. Therefore, age-specific mortality rate provides more information that gross mortality rate. We will state a few examples.



The data were transferred via RExcel, similar to the previous cases and we plotted them similarly:

> plot(mu_muzi, type="l", main="Age- and gender-specific mortality rate", xlab="age group", ylab="mortality rate", col="blue",lty=4, ylim=c(0,252))

> par(new=TRUE)

> plot(mu_zeny, ,type="l", col="red",lty=3,, xlab="age group", ylab="mortality rate", ylim=c(0,252))

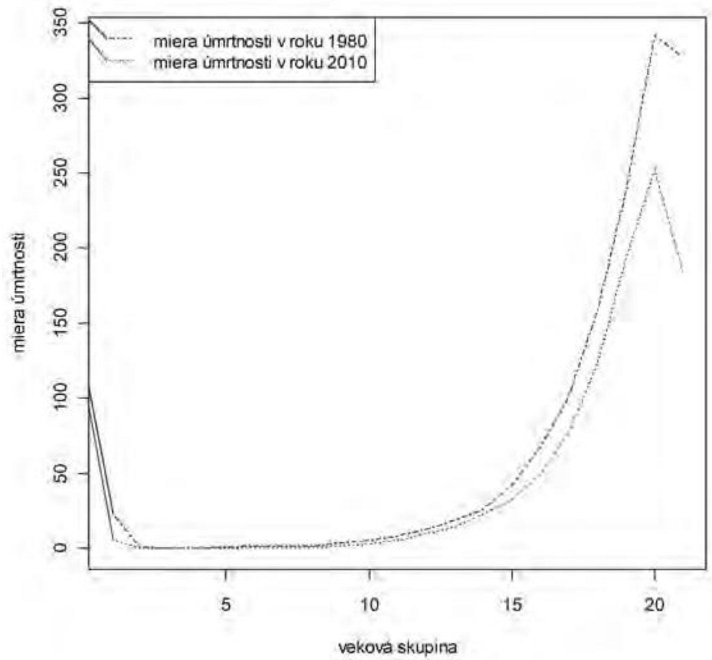
> par(new=TRUE)

> plot(mu_in total, type="l", col="green",lty=5, xlab="age group", ylab="mortality rate",ylim=c(0,252))> legend(„topleft", c(„men´s mortality rate","Women´s mortality rate", „mortality rate in total"),lty=c(4,3,5), col=c(„blue","red","green"))

Picture No. 8 Mortality rate according to age groups and gender in Slovakia in 2010. Source: POPIN SR

In the first one, we will illustrate age-specific mortality rate in Slovakia separately per both genders in 2010. The data come from DB *Slovak POPIN*, provided by Infostat.[27] On the Picture No. 8, we can see that actually all three curves are of shape U with mortality rate higher in the first two groups, then it drops and keeps low till the

[27] http://www.infostat.sk/slovakpopin/



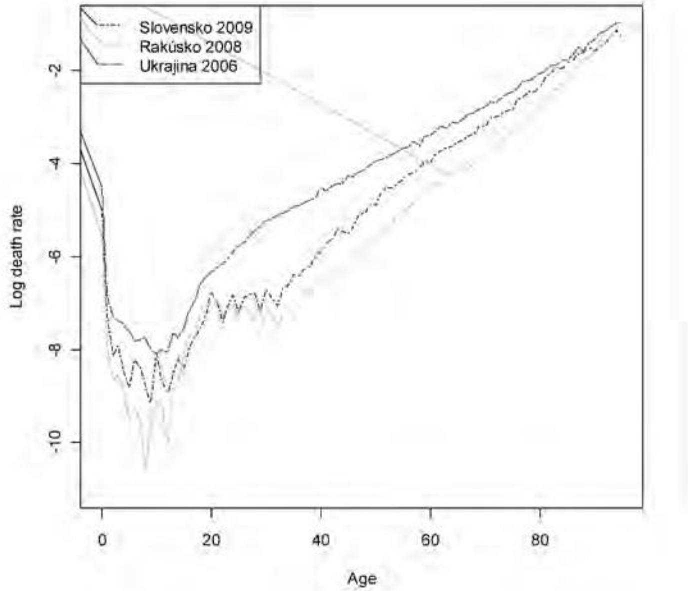
Picture No. 9 Men´s age-specific mortality rate in 5-year age groups during years 1980 and 2010 in Slovakia
Source: POPIN/Slovakia

middle age when it starts to rise. The maximum refers to approx. 80 years of age. Men reach higher value in the sharp rising curve section in all age groups.

To compare, look at the situation of men in Slovakia in years 1980 and 2010 (Picture No. 9). We can see much higher mortality rate in the first years of life. The difference gradually increases in younger adult age but is most significant mainly after the 50th year of age. Accordingly, we can think about the causes of such effect. These definitely include healthcare progress, availability of healthcare services (both preventive and therapeutic), as well as complex improvement of the Slovak population life quality.

Similarly, we can look at the age-specific mortality rate in three neighboring countries. We chose Austria as a country with the best economy, Ukraine that lags behind Slovakia in the terms of economy, and of course Slovakia. We drew the data from DB Human Mortality Database,[28] where the countries share the mortality related data. Project {R} provides for library (you can install it similar to other libraries in the project environment), which simplifies access to the data and offers functions for

[28] The Human Mortality Database http://www.mortality.org/



Script for age-specific mortality rate comparison in 3 countries: Slovakia, Austria and Ukraine
Data reading from Mortality database through demography library call. Data storage in the variable SK
> *SK <- read.demogdata(„Mx_1x1_SK.txt“,type=“mortality“, label=“Slovakia“)*
select data for 2009
> *SK.2009 <- extract.years(SK,2009)*
> *SK.2009*
Reading data for Ukraine
> *UI <- read.demogdata(„Mx_1x1_UI.txt“,type=“mortality“, label=“Ukraine“)*
> *UI.2006 <- extract.years(UI,2006)*
> *UI.2006*
> Reading data for Austria
> *A <- read.demogdata(„Mx_1x1_A.txt“,type=“mortality“, label=“Austria“)*
> *A_2008 <- extract.years(A,2008)*
> *A_2008*
Graphical illustration of mesaures for particular countries
parameter max.age=95 chooses only age grups up to
95 years of age parameter series=“male“ selects only men´s data
> *plot(SK.2009, main=“Mortality rate Slovakia 2009, Austria 2008, Ukraine 2006, men“, max. age=95, series=“male“, col=“blue“,lty=4, ylim=c(-11,-1))*
par(new=TRUE)
> *plot(A_2008, series=“male“, main=““, max.age=95, col=“green“,lty=5, ylim=c(-11,-1)) par(new=TRUE)*
> *plot(UI.2006, series=“male“, main=““, max.age=95, col=“red“,lty=6,ylim=c(-11,-1))*
legend insertion
> *legend(„topleft“, c(„Slovakia 2009“,“Austria 2008“, „Ukraine 2006“),lty=c(4,5, 6), col=c(„blue“,“green“, „red“))*

Picture No. 10 Comparison of mortality rate in 3 countries: Slovakia, Austria and Ukraine

Source: The Human Mortality Database

processing, for example plotting in graphs. Resulting three curves (Picture No. 10) document the differences in men´s mortality according to age at the men who died in particular year. DB doesn´t work with age groups; if we wish to create some, we would face time demanding data arrangement in age groups. For our needs, particular years are sufficient. We can see apparent differences in particular countries. While Austria reports the lowest mortality rate in all age groups, Ukraine is on the opposite spectrum side. Slovakia is somewhere in between. Like before, we should consider the causes of such effect.

Age-specific life expectancy or death likelihood

Looking at the mortality rate calculation method, we can see that the rate can be interpreted as likelihood of dying in particular age. In compliance with the probability definition,[29] probability or likelihood of death in particular age interval is defined as number of deaths within the interval, divided by number of individuals living at the interval start. [27] Such concept is stated in the terms of the fact that the dying likelihood represents the basis of the life tables´ structure, and a major factor during the study of mortality development trends and comparison of surviving in various communities. The comparison doesn´t require standardization.

Along with the stated indicators, the Dictionary of Demographical Terms [6] contains further measures derived from the death counts, contained in the table No. 10.

Cause-based mortality is not mentioned in this chapter and it will be described in the chapter dealing with measurement of population burden by diseases.

Mortality rate application to health studies

Mortality rate also highlights health problems of an individual and population. In the environment where many risk factors affect human´s health and illness (lifestyle, economy, social factors, healthcare and other services), they mutually influence one another and further influence mortality rate that can be characterized as one of the indicators of such factors. Local and regional mortality rate differences represent a significant indicator of setting forth public health program priorities and evaluation of long-term effects of such programs.

On the other hand, interpreting such effects requires careful approach. People studying this subject matter should take in account certain limitations resulting from the factors that limit the mortality rate interpretation value.

[29] Classical definition pursuant to Pierre Simone de Laplace states that probability corresponds to the number of relevant cases vs number of all possible cases. In other words, proportion of the situations when that occurs what we are interested in to the sum of situations when that occurs what we are interested in and the situations when that what we are interested doesn´t occur.

Rate	Definition	Note
Child mortality, under five mortality	Children mortality usually within 1 – 4 years of age	Upper limit isn´t defined And has been moved, e.g. up to 5 years of age
Infant mortality	Mortality of children up to 1 year of age.	
Endogenous infant mortality	Infant mortality for endogenous Defects, delivery defects, etc.	
Exogenous infant mortality	Infant mortality – infectious, parasitary diseases, respiratory, digestive diseases, etc.	
Gross death rate	Total deaths to population median, usually per a year	Usually expressed in pro mil.
Maternal mortality	Maternal mortality together with Pregnancy, delivery and puerperium.	
Child death rate	Total died children of age 1 – 4 To the children median count of The same age group per year.	Usually per1000 citizens. Because of data availability, UNICEF Defines child death rate (under – 5 mortality rate) as dead kids count below 5 years of age per 1000 Live born kids.
Infant mortality rate (IMR)	Total dead kids up to 1 year of Age to total born live kids Usually per 1 year.	Usually expressed in pro mil.
Maternal mortality rate (MMR)	Total mothers who died in rel. To pregnancy, delivery or Perpuerium to the count of born Live kids per year	Usually expressed in pro mil.
Late neonatal mortality rate	Headcount of the dead in age 7-27 days to the count of born live Kids usually per a year	Usually expressed in pro mil.
Neonatal mortality rate	Headcount of the dead in age 0-27 days to the count of born live Kids usually per a year	Usually expressed in pro mil.
Prenatal mortality rate	Number of dead births and dead Up to 7 days to the count of live Births usually per a year.	Usually expressed in pro mil.
Post-neonatal mortality rate	Total dead children from 28th Day up to 1 year to the count of Live births per a year.	Usually expressed in pro mil.
Post partum mortality rate	Headcount of the dead in age 0-2 days to the count of live births Per a year.	Usually expressed in pro mil.

Early neonatal mortality rate	Headcount of the dead in age 0-6 days to the count of live birth Usually per a year.	Usually expressed in pro mil.
Cause-specific death rate	Total deaths with particular cause Of death to population median Per a year	Usually expressed in pro mil.
Age-specific mortality rate, age-specific death rate	Total deads of particular age to Population median of the same Age per a year.	Usually per 1000 persons. Calculated also separately Per genders. Known also as specific mortality rate
First day mortality rate	Total deads within the first 24 Hours to the count of live births Usually per a year.	Usually expressed in pro mil.
Late neonatal mortality	Children mortality within 7-27 days	
Neonatal mortality	Children mortality within days 0-27 - during the first 4 weeks.	
Prenatal death	Mortality and dead natality rate Of kids within 7 days after delivery	Mortality around delivery, Except for mortality before abortion
Post-neonatal mortality	Children mortality from 28th day Up to 1 year of age.	
Post partum mortality	Children mortality within 0-2 Days – within first 3 days	
Early neonatal mortality	Children mortality within 0-6 Days – within first 6 days	
Mortality	Deaths in population monitored As a mass demographical phenomena	
Cause-specific mortality	Mortality classified according to Cause/ group of causes.	
First day mortality	Mortality rate during the first 24 hours of life.	

Table No. 10 Various mortality rate types used in demography

One of them refers to close correlation to age structure of studied population that should be always taken in account. So far, death is a natural life end even if its causes are undesirable. We shouldn´t forget data standardization and discrepancies resulting from death recording mainly according to death cause. Care should be taken when interpreting so called statistical significance in the way that neglects stochastic nature of such effects in the nature and society.

Demographical transition and health

We started our excursion in the history with Malthus who warned on unlimited population growth and suggested the methods how to avoid catastrophe. In the history, even the latest, we see efforts of intervention in the population development. China is generally known example thereof. In 1979, the Chinese Government implemented ambitious program of market reform following the economic stagnation as a consequence of the Cultural Revolution. The program included strict population regulation and restriction of having max. 1 child per a family, mainly in the cities. The strategy of a family with 1 child resulted in major discrepancies in the gender ratio and related problems in the society life and labor force. [28]

While the deeds from the era before, of regular birth and death recordkeeping on the national level is not satisfactory, health has improved enormously during the last three centuries without any doubts. It is likely that majority of all children died or was killed during major part of the human existence during the first years after delivery. More than 95% of children survive till the adult age in technologically developed countries.

Theory of demographical transition has been mostly recognized amongst those trying to explain the changes at reproduction behavior of population in well-developed countries for the last approx. 200 years. Mortality rate significantly decreased and population increased at the demographical transition start. Subsequently, natality rate was reduced and population was stabilized. **Demographic transition** represents a major change at population regime, characterized with significant decrease in mortality and natality rate. According to UN, it has been change from high to low mortality and natality rate as a result of industrialization and modernization. Demographical transition is sometimes called also the *First Demographical Transition*. [6] Many authors have made efforts to explain the causes leading to human life prolongation. Among other aspects, they highlighted also healthcare role. McKeown [29] sees the role of medicine in the broadest meaning in two areas: prevention of diseases and care of the ill.

The 2nd demographic transition has been characterized with changed reproduction behavior of the population, decreasing natality rate, change at family behaviors and shift in the value system. These changes reflect the post-modernism in the population development. This demographical transition is sometimes called also the *Second Demographical Transition* and has been reported since the end of the first demographic transition. Fertility has gradually dropped below two children per a woman, as well as marriage rate, increasing divorce rate and increasing age of the first marriage. Partnership without marriage has become common and there are ever more children born out of the marriage. Children are no more the family center of interest. All these effects resulted in gradual decrease of the population and its ageing. Slovakia follows the demographical transition with the delay of a few decades, compared to the most developed countries, and is a few decades ahead of the developing countries. Mortality rate sharply dropped and the mean life length increased in the 50s of the 20th century. At the time, it was common to marry at least once per life at 90 % men and women. From the 60s, difference in demographical development has been observed between the Western and Eastern Europe. At the time, high marriage and natality rate was reported in Slovakia thanks to traditional value orientation of the population and to social and habitation strategy focused on the support of families with children. Till the end of 80s, Slovakia had had one of the highest natality rates in Europe. Significant changes occurred in the 90s. Natality rate sharply decreased below two children per a woman, as well as marriage rate; average age of the first marriage increased as well as the count of partners living together without marriage and children born out of the marriage. Reduction of induced abortions by more than 60% represents a positive trend of the 90s. [30]

Summary

Demography represents a fundamental population science. It is dedicated to description of effects reported within population and seeks explanation thereof. Therefore, demography is interconnected with the sciences dedicated to health and illness of groups of people, as well as those dealing with social, economic and other development of the analyzed society. For such reasons, many indicators considered primarily demographical are used in the study of population groups in the terms of health and illness. We need to be knowledgeable of the demographical methods mainly when solving the issues of public health, preparing new programs or projects and evaluating their efficiency, and when estimating the development of human resources and human needs; in order to be able to apply these methods properly and interpret their results in combination with methods of the public health study. In this chapter, we discussed fundamental terms and methods used in the demography. However, more profound study exceeds the publication intents. Reader who has become interested in this science can find many sources of further study. We stated also the examples of demographical data processing in the program environment{R} and its interconnection with Excel in the RExcel environment. Both methods shall be used also in the following chapters.

References

1. Source: Statistical Office SR. *Demographics*. 2010 [cited 2010 28. 08.]; Available from: http://portal.statis-tics.sk/showdoc.do?docid=2434.
2. Kalibová, K. *Introduction to Demographics*. 2 ed. 2001, University Karlova, Prague, Karolinum.
3. Porter, D. *Health, Civilization, and the State: A History of Public Health from Ancient to Modern Times*. Welcome Institute Series in the History of Medicine Series. 1999: Routledge.
4. Porta, M., Greenland, S., Last, J. M. Editors. *Dictionary of Epidemiology*. 2008, Oxford University

Press, New York: Oxford. 299.
5. Rusnák, M., Rusnáková, V., Majdan, M. *Bio-Statistics for Public Healthcare Students,* I. ed, 2010, Trnava: Typi Universitatis Tyrnaviensis, Trnava University publishing agency. 216.
6. Jurčová, D. *Dictionary of Demographical Terms*. 2005, INFOstat – INSTITUTE OF INFORMATICS AND STATISTICS. Research Center of Demographics: Bratislava.
7. Source: Statistical Office. *System of demographical statistics and foreign migration statistics in Slovakia*. 2011 [cited 2011 25.2.]; Available from: http://portal.statistics.sk/.
8. Sirimi, N., Goulis, D. G. *Obesity in pregnancy.* Hormones (Athens), 2010. **9**(4): p. 299 – 306.
9. Ozdemir, S. et al., *Assessment of ischemia-modified albumin level in patients with recurrent pregnancy loss during the first trimester.* Eur J Obstet Gynecol Reprod Biol, 2010.
10. Burri, A. V., Cherkas, L., Spector, T. D. *Exploring genetic and environmental influences on miscarriage rates: a twin study.* Twin Res Hum Genet, 2010. **13**(2): p. 201 – 206.
11. Dubiella-Jackowska, A., Polkowska, Z. ,Namiennik, J. *Platinum group elements in the environment: emissions and exposure.* Rev Environ Contam Toxicol, 2009. **199**: p. 111 – 135.
12. Agyemang, C. et al. *The effect of neighborhood income and deprivation on pregnancy outcomes in Amsterdam, The Netherlands.* J Epidemiol Community Health, 2009. **63**(9): p. 755 – 760.
13. Regan, L., Braude, P. R., Trembath, P. L. *Influence of past reproductive performance on risk of spontaneous abortion.* BMJ, 1989. **299**(6698): p. 541 – 545.
14. Mosquera Tenreiro, C. et al. *Prevalence and secular trend of congenital defects in Asturias, Spain. The need for clinical-epidemiological surveillance.* Gac Sanit, 2009. **23**(4): p. 300 – 305.
15. Konje, J. C., Obisesan, K. A., Ladipo, O. A. *Health and economic consequences of septic induced abortion.* Int J Gynaecol Obstet, 1992. **37**(3): p. 193 – 197.
16. Goldie, S. J. et al. *Alternative strategies to reduce maternal mortality in India: a cost-effectiveness analysis.* PLoS Med, 2010. **7**(4): p. e1000264.
17. Davis, A. R., Beasley, A. D. *Abortion in adolescents: epidemiology, confidentiality, and methods.* Curr Opin Obstet Gynecol, 2009. **21**(5): p. 390 – 395.
18. Hodorogea, S., Comendant, R. *Prevention of unsafe abortion in countries of Central Eastern Europe and Central Asia.* Int J Gynaecol Obstet, 2010. **110 Suppl**: p. 34 – 37.
19. Medoff, M. H. *Non-marital births and state abortion policies.* Soc Work Public Health, 2010. **25**(5): p. 454 – 469.
20. Myers, J. E., Seif, M. W. *Global perspective of legal abortion – Trends analysis and accessibility.* Best Pract Res Clin Obstet Gynaecol, 2010. **24**(4): p. 457 – 466.
21. Coles, M. S. et al. *How are restrictive abortion statutes associated with unintended teen birth?* J Ado-lesc Health, 2010. **47**(2): p. 160 – 167.
22. NCZI. *Potraty v SR 2009. Abortions in the Slovak Republic 2009*. 2010, National Center of Healthcare Information: Bratislava.
23. Smith, M. P., Olatunde, O., White, C. *Monitoring inequalities in health expectancies in England – small area analyses from the Census 2001 and General Household Survey 2001 – 2005.* Health Stat Q, 2010(46): p. 51 – 68.
24. Popham, F., Bambra, C. *Evidence from the 2001 English Census on the contribution of employment status to the social gradient in self-rated health.* J Epidemiol Community Health, 2010. **64**(3): p. 277 – 280.
25. EUROSTAT. *Health statistics – Atlas on mortality in the European Union*. Statistical books. 2009, Luxembourg: Office for Official Publications of the European Communities. 215 p.
26. Shryock, H. S., Siegel, J. S. et al. *The Methods and Materials of Demography*. 4 ed. Vol. 2. 1980 Washington Dc Us Bureau of the Census, SfegcT and Associates fourth Printing rev Us Government Printing Office.

27. Chiang, C. L. *The Life Table and its Applications*. 1984, MAlaBAR,FLORIDA: Robert E. Krieger Publishing Company. 289.
28. Ding, Q. J., Hesketh, T. *Family size, fertility preferences, and sex ratio in China in the era of the one child family policy: results from national family planning and reproductive health survey*. BMJ, 2006. **333**(7564): p. 371 – 373.
29. McKeown, T. *The Role of Medicine*. 1976: The Nuffield Provincial Hospital Trust.
30. Vaňo, B., Jurčová, D., Mészáros, J. *Introduction to Demographics*. 2003: Civil Association Social Work.

Chapter 8

Burden of population illness

Chapter objectives

The previous chapters were dedicated to definition of health and illness in the terms of major human life events: birth, death and life. Illness consideration is more complicated in many aspects. We encounter problems when trying to define disease, determine its onset and end; process of some diseases is complicated because of accompanying disorders; many disorders are hard to define or diagnose and there are cases with ambiguous classification. In such cases, experts should evaluate the Burden of Population Illness since they understand subtleties of some diseases. In this chapter, we will present the problems associated with definition of some disease types; then we look at the mortality according to death causes and finally we will present basic and deduced indicators of the Burden of Population Illness. Similar to the previous chapters, we will illustrate particular issues with examples taken from domestic and foreign sources.

Health and illness and their impact on life statistics

It is common in the medicine and health-related sciences that every expert has its own view of the same situation. In case of health and illness, an epidemiologist mainly deals with onset and spread of disease in the population while a clinician is mostly interested in the disease progress at the patient.

Experts in the area of social medicine have searched for definition of illness that would cover majority of uncertainties that we encounter nowadays. Healthcare statistician seeks the definition of a disease onset and end in order to record the case. For purposes hereof, we will focus on two areas: disease onset and end.

Health

What is health? Mankind has been seeking for the answer for ages. Health is condition when an individual is not ill. But when is a man completely healthy? At birth? When he feels good? When he doesn´t miss anything? We know that answers to such and similar questions are not easy at all. WHO made efforts to define health after the WWII. end.

Health is condition of complete physical, psychical and social well-being instead of only absence of weakness and illness.

Definition 1 Health according to WHO [1]

Since its introduction, definition of the health has been both used and criticized. Definitions not based on the idea of positive health are acceptable from conceptual point of view but often considered not functional enough, being too universal. [2] Contrary to the positive definition of health, majority of evaluations of health condition measurement is negative, in other words diseases and their consequences are measured instead of health. [3] With regard to public health, we are focused on epidemic, acute and chronic diseases, their causes and risk factors and the occurrence in population. We try to foresee diseases, reduce their consequences and enhance health. Those controlling healthcare provision and creating healthcare strategy have different angle of view. They are rather interested in population well-being, illness cost, and evaluation of inequalities and definition of priorities. Epidemiologists are oriented to human diseases, environment and social factors that can influence disease progress and distribution through changed frequency, and to define biological and ecological factors of diseases. [4]

Causality

Illness definition as stated above doesn´t automatically explain all situations that a human being can encounter. Many philosophers and physicians had dealt with the subject matter without any single conclusion made. [5] What should be considered standard and what exceeds this condition and refers to an illness – this is one of many discussed problems.

What is standard?

For example, if we consider common behavior as a criterion of standard, human nature, then we should consider a newborn feeding with milk from the bottle non-standard. Thus, would we consider a mother with not enough breast milk sick? Another example refers to agreed standards. Regular man in our environment is 160- 200 cm tall. Should then a shorter or taller man be considered sick? Many physiological parameters change with age; it is regular that hair pigment diminishes with age and a man turns gray. Is a gray head an illness? Sensitive ear cells function also diminishes with age and if a child can hear high sound frequency not heard by an adolescent, not speaking of an adult or elderly man…. . Thus, is it a symptom of illness? If not, what about high blood pressure if we know that it rises with age as well? For such reasons, let´s adopt standard that we consider appropriate and interpret illness as deviation from the standard limits. We would like to point out to the uncertainties that bother professional public but to provide a complete philosophy of what is and isn´t standard.

Causality

Situation is even more complicated if we start answering question about the cause of particular illness. A non-professional could say that the answer is easy since we know the causes of diseases – tuberculosis is caused by bacteria, influenza is caused by virus and smoking causes lung cancer. Anyway, what if a man encounters the Koch bacillus but he isn´t infected by tuberculosis? Or what about a smoker who smokes for the whole life but dies of other cause than lung cancer? In 1890, Robert Koch formulated the principles of causality for infectious diseases:

1. Organism should be found in all ill persons and in no one who is healthy and its location should correspond to the observed disorder.
2. Organism shall not appear in any other illnesses as a random or pathogenetic parasite.
3. Organism can be cultivated out of the host body and repeatedly in a clear culture; such isolated organism should induce the same illness in other sensitive animal. [6]

However, validity of these postulates was undermined by new discoveries. It was found out that also healthy people can communicate some diseases, for example those caused by salmonella. Presence of a microorganism in heart wasn´t proved in case of diphtheria. As well, leprosy bacillus has never been cultivated but its presence was confirmed. Later discovery of viruses resulted in modification of the Koch´s postulates. In 1937, Rivers formulated the modification so as the postulates apply also to virus characteristics known at the time. He stated that a specific virus associated with an illness with certain degree of regularity should be discovered, confirmed in ill individuals not as a random finding but the cause of studies disease. The third Koch´s postulate cannot be applied since viruses require live culture for cultivation.

The situation got even more complicated upon massive onset of chronic, non-infectious diseases. Unambiguous causality could even hardly be proved here and it was the start of period when the causality has been expressed through probability (stochastic). Methods of causal relations seeking have evolved as well, mainly ever more complicated structure of studies and statistical methods. This situation is reflected in the Evans´s definition of causality, grouping the factors in 9 points (Table No. 1). [7]

1. Prevalence of disease should be significantly higher at those exposed to alleged causes than in case of control/ those not exposed.
2. Exposure to alleged cause should be more present at patients suffering the disease than in case of control/ those not exposed, provided that all risk factors are the same.
3. Disease incidence should be significantly higher at people exposed to alleged causes than those not exposed, as demonstrated by perspective studies.
4. Time-wise, diseases should follow the exposure to the alleged agens with bell-shaped incubation time distribution.
5. Host response spectrum should follow the exposure to alleged agens according to logical biological gradient from moderate to heavy.
6. Measurable host response following the exposure to alleged agens should regularly appear at those without the cause (i.e. anti-agens, cancer cells) or their size should increase if present before the exposure, contrary to people who weren´t exposed.
7. Experimental reproduction of disease should result in higher occurrence at animals or people adequately exposed to alleged cause, compared to those not exposed; the exposure can be deliberate at volunteers, experimentally induced in the lab or confirmed through controlled regulation of natural exposure.
8. Removal or change of alleged cause or transmission vector should reduce the disease occurrence (contaminated water or smoke control, removal of particular agens).
9. Prevention or change of host response to exposure to alleged cause should diminish or suppress disease (vaccination, cholesterol reduction medication, specific lymphocyte factor of cancer transmission).

Table No. 1 Criterions of the Evans´s definition of causality [7]

Causality should be strongly embedded in the human population health statistics and it has been further developed concurrently with the development of cause recognition technology. Its cognition represents an important part of the interpretation of population infectious and non-infectious illness burden. Epidemiology deals with identification of causality and Statistics of Health provides for such conscience to the professional public on regular basis.

Thus, we have to state that in spite of all, Statistics of Health is mainly the statistics of diseases. We regularly encounter people with deviation from full health condition. There are rare cases when we organize opinion poll, asking people on subjective feeling of health. Then we can obtain statistical data of people with full-health feeling, which however doesn´t mean they are completely healthy. Thus, should we talk about the statistics of the ill and the dead (even if one can die at full health, for example as a result of injury)? Therefore, basic categories are sufficient to consider for statistical purposes; since we are able to determine and recognize them with sufficient accuracy.

Lately, the interest in the statistics has increased since the new indicators describe both condition and the ways of change/ improvement. For example, indicator *Life Potential Loss* (PYLL) has been ever more used during evaluation of the population intervention results. Increasing interest in health variations and healthcare planning is also an important actor.

Cause-specific mortality

We talked about the death statistics in the chapter dedicated to demography where we deliberately excluded death-cause based classification from. Thus we included it in this chapter since it deals with population diseases. Cause-specific mortality classifies death according to causes. The term *Cause-specific death rate* could be used as well according to the Dictionary of Statistics. It is calculated as headcount of the dead per particular death cause to the population median, usually per a year and expressed in pro mil. [8] in the following formula:

$$M = \frac{D}{P} \times 10^5$$

Letter M refers to cause-specific mortality; D refers to headcount of the dead per particular death cause and P refers to the population median in particular year.

Headcount of the dead per main death cause provides information on the population health condition. The information can be used for calculation of total and premature mortality and serve as the basis for deduction of healthcare and public health strategy. Evaluation of efficiency of health intervention programs as well as vaccination and screening can take in account also cause-specific mortality. The measure is also important for identification of population risk groups. Mortality in case of certain specific diseases, for example malignant tumors, can be applied to evaluation of treatment efficiency or prevention of risk behaviors.

It is important to recognize the limitations resulting from cause-specific mortality recording method. In the chapter dedicated to demography, we included the *Dead Obduction Sheet* and Statistical Death Report, where further, secondary diseases are entered in along with primary death cause. Unfortunately, official statistics records only primary cause, ignoring those secondary ones. This is a problem; since cardiovascular system failure used to be entered as a primary death cause despite of being in fact a complication of the chronic disease. Looking at the mortality rate development in case of high blood pressure as a cause, we cannot say that statistical data truly reflect the situation in the fatal complications of hypertension. This serious deficiency of the cause-specific mortality rate meaning value is partly compensated with the fact that the records specify deaths caused by cardiovascular diseases.

Multiple death causes

The above stated reasons resulted in established analyses of multiple death causes. For example, the Australian Statistical Office has published multiple death causes since 1997. [9] This system enables provision of complex information on the group of diseases or their causes that contribute to fatal outcome but are not stated as a main death cause (e.g. Alzheimer´s disease, diabetes, pneumonia). It also proves the importance of correlation between the death causes in case of multiple diseases, e.g. investigates correlation between diabetes and heart ischemia. This approach enables closer information on the complexity of correlations between particular death causes, for example the nature of injury in relation to external causes.

Multiple death causes can be used for country-wise comparisons. For example, comparison of the data from Italy and France resulted in the observation that using this approach doesn´t change serious nature of diseases in the population. [10] Nevertheless, the role of background diseases has been disclosed, i.e. those that are not classified as direct death cause, for example hypertension. Similar results were obtained in USA. [11] Treatment of blood disorders, e.g. anemia or blood coagulation disorders can represent another cause of different results of mortality rate. These disorders can be caused by other diseases, e.g. liver disease or cancer therapy. Decubiti represent another example. Septicemia and other serious infectious diseases can contribute to the mortality picture. Accordingly, we see that knowledge of secondary death causes contributes to identification of a complex view of the health condition, hospitalization and their outcomes.

Disease occurrence

Case definition

A condition could appear in a person´s life when he feels that something´s gone wrong. He has fever, pain, feels tired, suffers nausea, puts weight on or off, doesn´t urinate enough or urinates too frequently, finds an increasing node upon palpation. Then he visits a physician who states a disease. The disease/ suspect disease is entered in the record and the physician starts with therapy or continues in examination. There are many diseases accompanied with warning symptoms, e.g. high blood pressure. It is identified by chance or during planned preventive examination. All in all, a disease appeared from statistical point of view.

Recording of confirmed disease or suspect disease represents a case from statistical point of view. In general, there is statistics of new diseases that appear within population in particular time and location. Since most diseases don´t have exact onset, we have to specify when the suspect condition can be called disease. Such definition is called *case definition*.

Case definition represents a common agreement on specification of certain case a disease of particular name. The definitions can differ according to disease type and diagnostic options. As an example, let´s talk about tuberculosis caused by Bacillus Koch. A disease used to be declared tuberculosis if we prove the presence of the pathogens. But sometimes there isn´t basic equipment available and we have to rely upon clinical symptoms only. Thus, we will declare tuberculosis a disease upon confirmed presence of Bacillus Koch in our conditions. This is the basis of case definition. As soon as the bacterium presence is confirmed, it is tuberculosis and a new case if it is the first case captured at particular person. It definitely doesn´t mean that the person suffers TBC since the moment of confirmed diagnosis. Disease symptoms could have appeared slowly together with organism struggling with it. The moment of confirmed diagnosis is a moment of new disease occurrence and when reported, the statistics of new TBS diseases will increase by one case.

There are places and situations when such laboratory proof is not available. Would we wait till the diagnosis confirmation? No, we wouldn´t , in case of persistent cough and frequent TBC occurrence, upon presence of further symptoms, we will declare suspect TBC occurrence and start with therapy. However, in such situation, the case definition doesn´t contain laboratory proof request.

Case definitions represent an important part of the disease surveillance systems. If the definitions are generally available and defined in particular environment, their application enhances the quality of surveillance, early warning and response systems. In case of faulty definitions, these don´t reflect real conditions of environment and surveillance interpretation value is doubtful.

Decision was adopted, applicable to infectious diseases subject to mandatory reporting within EU [12], which contains definitions of these diseases. The disease definitions are aimed at facilitating reporting of diseases and specific health issues. Clinical criterions should include general and relevant disease symptoms and signs that separately or jointly represent clear or preliminary clinical picture of the disease. Clinical criterions provide general characteristics of a disease and don´t have to indicate all indications required for clinical diagnosis of the disease. Laboratory criterions should represent the list of one or more laboratory methods applied to confirmation of the disease case. Epidemiologic criterions are considered met if epidemiological correlations can be confirmed. These correlations include man-to-man transmission, animal-to-man transmission, exposure to the effects of common source or contaminated food/ drinking water, environmental or laboratory exposure. Disease cases are classified as „possible", „probable" and „confirmed". The list of such defined diseases is contained in the tables No. 2 and 3.

Disease definition is more complicated in case of chronic diseases since the disease onset specification is practically impossible. Majority of these diseases continuously

---

Acquired immunodeficiency syndrome (Aids) and human immunodeficiency virus infection (HIv)
Anthrax (*Bacillus anthracis*)
Avian flu A/H5 or human A/H5N1
Botulism(*Clostridium botulinum*)
(*Streptococcus pneumoniae*)
Brucelosis (*Brucella spp.*)
Campylobacteriosis (*Campylobacter spp.*)
Chlamydia infection (*Chlamydia trachomatis*) incl. Lymphogranuloma venereum (LGV) Cholera (*Vibrio cholerae*)
Variant of Creutzfeldt-Jakob disease (VCJD) Cryptosporidiosis (*Cryptosporidium spp.*) Diphteria (*Corynebacterium diphtheriae and Corynebacterium ulcerans*)
Echinococciosis (*Echinococcus spp*)
Infections Escherichia Coli producing toxine SHIGA/VERO (Stec/VTEC)
Ghiardiasis (*Giardia lamblia*)
Ghonorrhoea ((*Neisseria gonorrhoeae*)
Haemophilic invasive meningitis
Haemophilus influenzae
Hepatitis A (*Hepatitis A Virus*) Akútna hepatitis B (*Hepatitis B virus*)
Hepatitída C (*Hepatitis C Virus*) Flu (*Influenza Virus*) Legionary disease (*Legionella spp.*) Leptospirosis (*Leptospira interrogans*) Listeriosis (*Listeria monocytogenes*) Mallaria (*Plasmodium spp.*)
Measles (*measles virus*)
Invasive meningococcus disease (*Neisseria meningitidis)*
Mumps (*Mumps virus*)
Pertussis (Black cough) (*Bordetella pertussis*) Pest (*Yersinia pestis*)
Pneumococcus invasive disease (-A)
Poliomyelitis (*Polio virus*)
Q fever (*Coxiella burnetii*)
Rabies (*Lyssa virus*) Rubella (*Rubella virus*)
Congenital rubella (incl. congenital rubella syndrome)
Salmonelosis (*Salmonella spp*. save S. Typhi and S. Paratyphi)
Serious acute respiratory syndrome– SARS (*SARS-coronavirus, SARS-CoV*)
Shigellosis (*Shigella spp.*)
Variola (*Variola virus*)
Syphillis (*Treponema pallidum*)
Syphilis, congenital and neonatory (*Treponema pallidum*)
Tetanus (*Clostridium tetani*)
Congenital toxoplasmosis (*Toxoplasma gondii*) Trichinellosis (*Trichinella spp.*)
Tuberculosis (*Mycobacterium tuberculosis complex*)
Tularaemia (*Francisella tularensis*) Typhus/ paratyphus fever (*Salmonella Typhi/ Paratyphi*)
Viral haemorrhagic fever
West Nile virus infection – WNV)
Yellow fever (*yellow fever virus*)
Yersiniosis (*Yersinia enterocolitica, Yersinia pseudo-tuberculosis*)

Table No. 2 Diseases defined in the EU Decree

AnthraX (*Bacillus anthracis*)
Clinical criterions:
Any person with at least one of the following clinical forms:
Skin form
At least one of the following symptoms:
– papular or vesicular lesion
– black eschar surrounded by edema
Intestinal form
– fever or chills, and
At least one of the following symptoms:
– strong abdominal pain
– diarrhea
Inhalation form
– fever or chills, and
At least one of the following symptoms:
– acute respiratory shortness syndrome
– radiological proof of mediastine expansion
Meningeal/ meningoencephalitic form:
– fever
At least one of the following symptoms:
– spasms
– unconsciousness
– meningeal symptoms
Septic form:
Laboratory criterions
– Isolation of *Bacillus anthracis* from clinical sample
– Identification of *Bacillus anthracis* – nucleic acid in clinical sample
Positive nasal swab result without clinical symptoms doesn´t contribute to confirmed diagnosis
Epidemiologic criterions
At least one of the following epidemiologic correlations:
– animal-to-man transmission:
– exposure to common source effects
– exposure to contaminated food/ drinking water effects
Disease case classification
A. Possible case – N/A
B. Probable case
Every person meeting clinical criterions being in epidemiologic correlation
C. Confirmed case
Every person meeting clinical and laboratory criterions

Table No. 3 Definition of anthrax case pursuant to [12] passes from the risk stage through the preliminary symptoms stage up to clinically manifested disease outbreak. On the hypertension example, we will demonstrate complicated nature of the situation. At first, let´s repeat that blood pressure is a parameter directly proportional to natural ageing of a person. It was proved many times that the first vascular changes appear in early adult age. [13–15] Therefore, the guidelines published by WHO [16] define hypertension and the start of therapy through systolic and diastolic blood pressure figures. It doesn´t mean that the disease started on the day when the one´s blood pressure reached values defined in the publication. It is an agreement supported by the studies about the efficiency of hypertension complications prevention at the adult.

Agreements about definition of majority of recognized diseases were created by national or international professional companies. For example definition of diabetes mellitus was created by WHO [17] and determines which clinical and laboratory indicators are required for confirmation of the diagnosis.

Why is it important to take in account these facts? It is because of health condition evaluation. Case definition doesn´t have to be only clinical but derived from hospitalization or visit of a physician. In the Canadian study about diabetes occurrence, diabetes mellitus case definition was used in the medical record in the terms "at least one hospitalization with diagnosis "diabetes" or at least two visits of physician with the same diagnosis within two years ". [18] If we are able to sufficiently define condition or disease, then we can better understand correlations of disease occurrence within population. We use two basic measures when describing a disease occurrence within population: incidence and prevalence.

Incidence

Incidence refers to new disease occurrence rate. It is expressed per population within which it occurred and the time of observance. Because of being named "rate", let´s remind how we define it: *Rate* represents number of demographical events of certain type in particular population group (mostly it is the middle class) per certain time period. [8] This is a demographical definition of rate but the same applies for epidemiology. Thus, a fraction numerator represents the number of new cases. They can be diseases but also other health-related events. In case of newly awarded disablement status, it is a disablement incidence, in case of death, we talk about mortality, despite of an incidence variant. It is important to recognize that we are interested in changed status in case of incidence, i.e. change from health condition, or the condition of not qualified/ identified problem, to the status when preconditions contained in the case definition are met. Thus, as soon as a person suffering diarrhea visits a physician and he suspects salmonella-caused disease, the case is recorded as salmonella infection on the day of the physician´s visit. However, the disease onset could have been dated much earlier. It is even easier in case of diabetes. Until a person visits a specialist, he can suffer condition classified as diabetes according to the definition but doesn´t bear the disease signs. As soon as the condition is confirmed by physician, the condition change is recorded and the disease case rate rises by 1. Let´s look closer at the fraction numerator. We say that it refers to headcount of citizens endangered by the condition. Thus, the numerator should include only individuals without the disease and concurrently endangered by it. Thus, women after hysterectomy shouldn´t be included in the calculation of cervical cancer incidence.

We haven´t stated a typical disease incidence formula yet. Let´s mark it with *I*, then the new cases number shall be *N*, endangered population shall refer to *Pr* (in order to distinguish usual probability sign).

$$I = \frac{N}{Pr} \times 10^{n}$$

You might be surprised by the absence of usual multiplication by ten thsd. in the formula. It has many reasons; one of them is that the multiplication is only used to modify the result to be as simple as possible, for example to avoid too many zeroes after decimal point. Therefore, letter *n* represents number 1 -5, i.e. $10^1 = 10$, $10^2 = 100$, $10^3 = 1\ 000$, $10^4 = 10\ 000$, $10^5 = 100\ 000$ etc..

We distinguish two incidence types: *cumulative incidence and incidence rate*.

Cumulative incidence

Cumulative incidence (or other expressions used: *incidence proportion, attack rate, risk, disease occurrence probability*) results from the precondition that we are able to monitor the whole population exposed to risk since the start up to the end of monitoring. We can apply cumulative incidence in practical epidemiology and cohort studies. For example, epidemiologist examines diarrhea epidemic outbreak at wedding party guests. If there were 60 guests at the party and 18 of them were affected by diarrhea within 24 hours, then the diarrhea cumulative incidence refers to:

$$I = \frac{18}{60} \times 100 = 30$$

Thus, there were 30% new cases or 30% risk rate.

It is important to take in account that the monitoring time must be clearly specified. In the example, we mentioned 24 hours. Had we prolong the monitoring to 24 days, then we wouldn´t measure the condition incidence in relation to the poisoning during the wedding party. In this way, time becomes a major factor when considering incidence, mainly if persons are not exposed to risk for equally long time. Somebody is exposed to risk earlier and somebody later; somebody is excluded from monitoring for other reasons. In such case, if the loss is proportionally huge, it is necessary to apply some adjustments as in case of the survival studies.

Cumulative incidence obtained from correct method represents the disease occurrence risk and can be expressed as probability. It is a direct estimation of relative risk.

Incidence rate/ density

We are not always able to exactly measure the monitoring start, evtl. duration of a person´s exposure to risk. For example, we have to rely upon the age of person with diagnosed disease during monitoring of mandatorily reported diseases. Thus, headcount of persons living in particular time is used as a fraction numerator. Such determined incidence is called *incidence density* and it is actual incidence rate. Let´s describe it better since it is used very frequently. At first, remind that the population headcount represents the fraction numerator (headcount of city, district or country population) as of July 01 of particular year, thus population median in the year. We can obtain it from the census or demographic estimation. It is an average population headcount exposed to risk with presumption that each member thereof is monitored for the same time.

On the example (Table No. 4) we present a study dedicated to infectious disease occurrence at a cohort of young women. We are interested in the draft study and method of interpretation of its results. At first, look at the definition of the young women´s population. They are women visiting gynecological or other outpatient´s department with the aim to pass examination or consultations. There are exact rules of inclusion in the study, also stating classification of the findings, what was considered a new case or persisting infection/ recurrence. Infection prevalence at the time of recruitment was determined according to entry examination findings. In the terms of incidence, new-case incidence and calculation of incidence compared to duration of the monitoring for the study purposes was the point of our interest. Thus, the incidence calculation is as follows:

$$I = \frac{\textit{New cases}}{\textit{Study duration}} \times 100 = \frac{47}{1056{,}34} \times 10 = 4{,}45$$

We can interpret the result as an incidence during 100 man-years; or in other words, if we monitored 100 women during one year, we would observe occurrence of new Chlamydia infection in 4 cases. We should analyze the stated man-years count. The authors state that 877 women attended the study for the whole 12 months. Thus, 239 of total 1116 women prematurely left the study and 877 women represented 877 man-years while 239 women attended 1056 – 877 = 179 man-years. That means, one woman contributed with approx. 0.75 year (9 months). From that, we can derive that the project was very successful and there were very few women who left.

Objective

The study was focused on the Chlamydia infection incidence estimation and repeated infections, and was aimed at examining the dynamics of organism burden by Chlamydia at prevalent, incidental and recurring infections at young Australian women.

Methodic

Women were included in the study upon visiting the outpatient´s department for any reason during the study, provided they had had vaginal sex with a man, weren´t pregnant, were able to understand English written text and could be contacted by post during 12-month taking study. The women were tested upon recruitment (basic test) and after 6 and 12 months. Women with positive test in any stage were retested after 3 months. All subsequent tests included collection of vaginal swabs shipped by mail. Data from total 1 116 women of age 16 – 25 who visited the primary case outpatient´s departments in Australia were collected.

Prevalent infection was defined as a positive Chlamydia presence diagnosed by subsequent test during the study. The women were classified like with infection recurrence in case of positive Chlamydia test repeatedly during the monitoring following the previous positive test or without negative test in between the positive test results. The following results were defined as infection recurrence.

Results

Total 1116 women attended the study; thereof 79% (877 women) remained in the study for the whole 12 months. Chlamydia was identified with prevalence 4,9% (95% CI: 3.7%, 6.4%) at the time of recruitment. During the monitoring, 47 cases of new infections were classified as infections of 1056.34 men-years of monitoring with the incidence rate 4.4 per 100 men-years (95% CI: 3.3, 5.9).

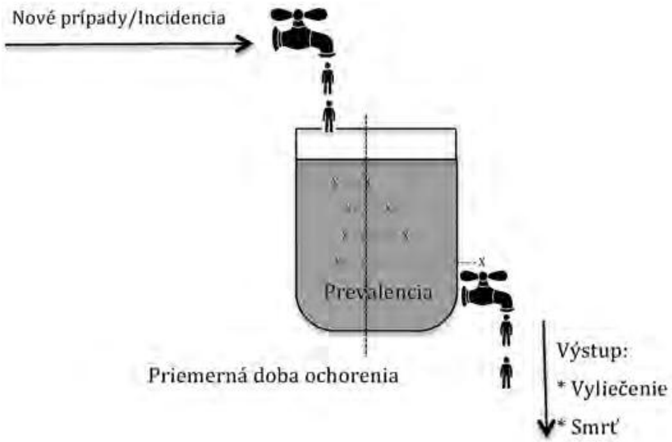
Table No.4 Example of incidence observation in population study [19]

Prevalence

Prevalence or illness rate can be interpreted as proportion of individuals in population with certain disease, syndrome or sign in particular time or during certain time period. If incidence refers to disease or disease condition increase rate, prevalence refers to total disease occurrence rate. Thus, contrary to incidence, disease rate is interested in new disease cases as well as all cases within limited time and location.

Prevalence can be displayed as a reservoir where new disease cases flow in, remain there within the disease duration and flow out only in case of healing or death (Picture No. 1). As seen on the picture, there is the same level in the reservoir (disease or syndrome) but individual response can differ, depending on applied time interval to prevalence identification.

Prevalence calculation method is similar to that of incidence, expressed with formula. Let´s mark prevalence with *Ch* (disease rate), all cases with *D*, population exposed to risk with *Pr* (in order to distinguish usual probability sign).

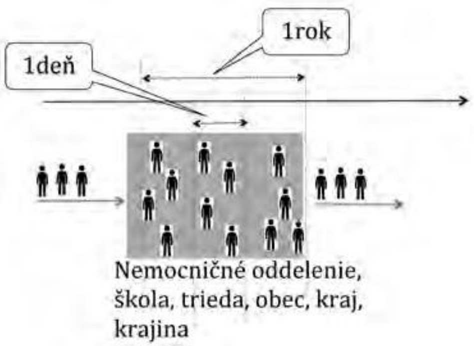


Picture No. 1 Relation between incidence and prevalence

$$Ch = \frac{D}{Pr} \times 10^{n}$$

To make the result better readable, we will add multiplication of ten at the end.

If we want to know disease prevalence in certain time, it is *point prevalence.* This is about the count of the ill that have been recognized in particular time and location. It is often used in the studies in well-defined populations, for example hospitalized patients. Disease prevalence is monitored also within certain time interval; in such case we talk about *period prevalence*. Reader can find further three terms in the Dictionary of Epidemiology [20] , each of them describing prevalence during certain period. *Annual prevalence* determined rate of an individual suffering the disease or any condition at any time during year. They are individuals who fell ill or acquired certain condition during the year regardless being present in particular territory before.



Picture No. 2 Point and period /annual prevalence

Those who will be healed during the year or die, or continue to be ill also in the following year/ years, shall be also counted in (Picture No. 2). Annual prevalence is used only rarely.

Another rarely used term is *lifetime prevalence.* It is the rate of individuals with present disease/ condition during life. Similar definition applies to *one-year prevalence;* it is the rate of people suffering certain disease or disorder during a calendar year. One should carefully decide on the value used in place of fraction numerator in each stated case.

Currently, point prevalence of hospital infections has become a routine method in many EU hospitals as proved by the published implementation guidelines. [21] Example of such study from Italy [22] is contained in the following table (Table No. 5). The study was conducted according to internationally agreed methodology and it guarantees certain quality level. Selection of population on which the study was

---

Study objective
Identify trends of prescriptions and correlation with antimicrobial resistance, identify improvement areas.
Methodic
The study was conducted in all hospitals´ wards within July 04 – 16, 2007. Antibiotics administering data were obtained from the records of all patients hospitalized for more than 48 hours. Age, gender, admission diagnosis, type and number of administered antibiotics were recorded at each child. Prescribed antimicrobial medication based on clinical symptoms but not confirmed micro-biologically (i.e. only empirically) were marked and compared to those with confirmed laboratory findings (based on microbiological findings), or those related to prophylaxis.
Prescription of antibiotics was calculated per entire hospital and ward type.
Results
Total 412 records of hospitalized children were processed. Antibiotics use prevalence was higher at older children (from 33.7 % at kids of age 0-6 months (32/95) up to 42.4 % at kids of age 7 mos. – 5 yrs. (61/144) and 49.1% at kids older than 5 years of age (85/173). Antibiotics use prevalence was 37.7 % at the wards with applied conservative therapy; 51.1% at the Emergency Room and 52.2 % at surgery wards. Of total 181 children treated by antibiotics, 78 (43.8 %) were administered with more than one medication. Thus, combined treatment prevalence referred to 18.9 %.
Discussion
Compared to the findings from other EU countries, our results demonstrated higher antibiotics use prevalence than in the Netherlands and Switzerland at the end of 1990 and 2000, (prevalence 36 %), but lower than in UK in 2006 (49 %). Prescription rate supported by microbiological findings was similar to the EU survey.

Table No. 5 Study of antibiotics administering point prevalence in hospital. [22]

Rate	Numerator	Denominator
Cumulative incidence (Attack Rate, Risk)	Total new disease cases per certain time internal	Population at the beginning Of time interval
Secondary Attack Rate	Total new disease cases Amongst contacts	Total number of contacts
Incidence rate/ density	Total new disease cases per Certain period	Total years of man´s observati-on or average population during monitoring interval
Point prevalence	Total concurrently monitored cases (new and existing) in Certain point	Population in the same time point
Period prevalence	Total cases (new and existing) During specific period	Average population/ median

Table No. 6 Most used disease prevalence rates pursuant to [23]

Prevalence increased by	**Prevalence decreased by**
• Longer disease duration	• Shorter disease duration
• Prolongation of life without therapy	• High mortality of specific disease
• Increased incidence (new cases)	• Reduced incidence
• New cases (migration)	• Healthy people (migration)
• Leave by healthy people (migration)	• Leave by ill people (migration)
• Disease sensitive people arrival	• Improved disease outcome
• Improved diagnostics (better reporting)	

Table No. 7 Processes affecting disease prevalence pursuant to [24]

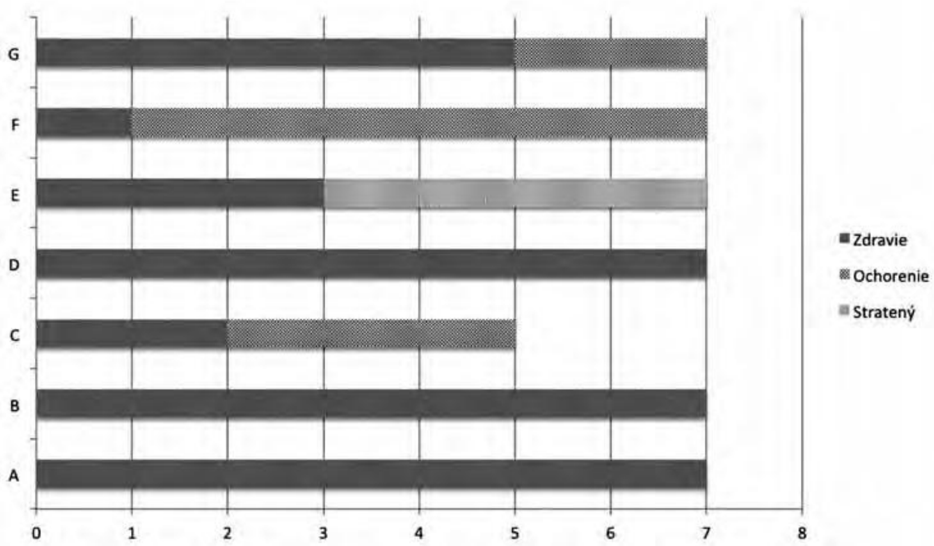
conducted was also properly defined – all people were hospitalized patients. The study time was strictly limited to 14 days. Thus, point prevalence doesn´t have to mean one day or one hour but we rather interpret it as a short time interval. One should notice that disease wasn´t the study subject but condition. It was defined by administering of antibiotics.

Relation between disease duration and complete healing can vary. Infectious diseases can take a few days and end with complete healing. If we know average disease duration, it is possible to estimate disease prevalence in particular period.

Prevalence can be affected by various factors, e.g. such that don´t directly relate to disease, for example migration (Table No. 7). Thus, prevalence effect on causality investigation is limited. Nevertheless, prevalence knowledge is important for healthcare service planning, allocation of resources and estimation of the needs.

Relation between measures

We will try to demonstrate relation between particular measures: incidence, prevalence, mortality rate in the terms of disease progress. Let´s use hypothetical population of 7 persons to be monitored for 7 years. During the monitoring, one of them dies, three persons fall ill and one is lost during monitoring. The scheme on the following picture (Picture No. 3) illustrates the relations.



Picture No. 3 Relations between disease occurrence rate pursuant to [24]

Persons A and B live healthy without any disease within entire monitoring, thus each of them has 7 healthy years. The third one, marked C, was not so lucky, he fell ill in the 2nd year of monitoring and died after three years. The person D had lived for 7 years without falling ill and person E disappeared from the monitoring after three years. The case F was healthy during the 1st year and then he fell ill and had been ill within entire monitoring; the case G fell ill in the 5th year of monitoring and had been ill within two following years.

What could we derive from the above facts? Let´s presume that all who fell ill suffered the same disease. Then the disease incidence will be calculated as the number of newly ill, i.e. 3, vs years exposed to risk, thus period in which the individuals were healthy (A – 7, B – 7, C – 2, D – 7, E – 3, F – 1, G – 5), which is in total 32 years. Accordingly, the incidence referred to 9.4 cases per 100 man-years in exposure to risk.

Cumulative incidence represents the number of ill, i.e. 3 vs all that were healthy at the start of monitoring, i.e. 7. Cumulative incidence will refer to 43/100 man-years in exposure to risk upon recalculation per 100 man-years.

Average duration of disease referred to the years lived in the disease, divided by the number of the ill, i.e. 11/3 = 3.7 years.

Prevalence depends on the moment when we make observations: prevalence refers to 0 in the 1st year, 2 of 6 living monitored persons are ill in the 4th year, i.e. prevalence refers to 33 cases per 100 men exposed to risk.

Summary

We scooped in the area of intersection of a few scientific disciplines, mainly demography and epidemiology. We demonstrated disease occurrence measuring methods in population and dealt also with problems resulting from variability of pathological process nature. Thus we have prepared the baseline for detail study of the processes and their causes. In the following chapters, we will describe indicators derived from those discussed in this chapter.

References

1. WHO. *Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States* 1948, WHO: New York. p. 100.
2. Basch, P. F.*Textbook of international health*. 1990, Oxford: Oxford University Press.
3. Bergner, M. *Measurement of health status.* Med Care, 1985. **23**(5): p. 696 – 704.
4. Sermet, C., Cambol, E. *Measuring the state of Health.* In: *Demography: Analysis and Synthesis, Four Volume Set, Volume 1–4: A Treatise in Population*, G. Caselli, G. Wunsch, and J. Vallin, Editors. 2005, Academic Press. 2976 pages
5. Koeslag, J. *What is normal?* S Afr Med J, 1993. **83**(1): p. 47 – 50.
6. Tan, S. Y., Berman, E. *Koch R (1843 – 1910): father of microbiology and Nobel laureate.* Singapore Med J, 2008. **49**(11): p. 854 – 855.
7. Evans AS. *Causation and Disease: The Henle-Koch Postulates Revisited.* THE YAle JOURNAL OF BIOLOGY AND MEDICINE, 1976. **49**: p. 175-195.
8. Jurčová, D. *Dictionary of Demographical Terms*. 2005, INFOstat – INSTITUTE OF INFORMATICS AND STATISTICS. Research Center of Demographics: Bratislava.
9. Australian Bureau of Statistics. *Multiple Cause of Death Analysis, 1997 – 2001*. 2003; Available from: http://www.abs.gov.au/ausstats/abs@.nsf/productsbytopic FDB92CC903BC3DC8CA2 56D6B0005A769?OpenDocument#.
10. Désesquelles A et al. *Revisiting the mortality of France and Italy with the multiple-cause-of-death*

approach. DEMOGRAPHIC RESEARCH, 2010. **23**(28): p. 771 – 806.

11. Manton, K. G., Stallard, E. *Temporal trends in U. S. multiple cause of death mortality data: 1968 to 1977.* Demography, 1982. **19**(4): p. 527 – 547.
12. EU *Decision dated Apr 28, 2008, amending Decision 2002/253/ES, establishing definitions of diseases for purpose of transmissible disease reporting in EC network based on decision of EU Parliament and EU Council No. 2119/98/ES.*, in *K (2008) 1589*, K.E. of EC, Editor. 2008: Brussels.
13. Enos, W. F., Holmes, R. H., Beyer, J. *Coronary disease among United States soldiers killed in action in Korea; preliminary report.* J Am Med Assoc, 1953. **152**(12): p. 1090 – 1093.
14. McNamara, J. J. et al.*Coronary artery disease in combat casualties in Vietnam.* JAMA, 1971. **216**(7): p. 1185 – 1187.
15. Solberg, L. A., Strong, J. P. *Risk factors and atherosclerotic lesions. A review of autopsy studies.* Arteriosclerosis, 1983. **3**(3): p. 187 – 198.
16. *2003 European Society of Hypertension-European Society of Cardiology guidelines for the management of arterial hypertension.* J Hypertens, 2003. **21**(6): p. 1011 – 1053.
17. *WHO. Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. R e p o r t of a WHO Consultation Part 1: Diagnosis and Classification of Diabetes Mellitus. 1999, World Health Organization, Department of Non-communicable Disease Surveillance: Geneva.*
18. *Diabetes in Canada: Highlights from the national Diabetes Surveillance System 2004 – 2005*. 2008, Her Majesty the Queen in Right Canada, represented by the Minister of Health. p. 19.
19. Walker, J. et al. *Chlamydia trachomatis Incidence and Re-Infection among Young Women – Behavioral and Microbiological Characteristics.* PLoS One, 2012. **7**(5): p. e37778.
20. Porta, M. *Dictionary of Epidemiology.* 2008, Oxford University Press, New York: Oxford. p. 289.
21. ECDC, *ECDC TECHNICAL DOCUMENT. Point prevalence survey of healthcare associated infections and antimicrobial use in European acute care hospitals – protocol version 4.3* 2012, European Centre for Disease Prevention and Control: Stockholm.
22. Ciofi Degli Atti, M. L. et al. *Point prevalence study of antibiotic use in a pediatric hospital in Italy.* Euro Surveill, 2008. **13**(41).
23. CDC. *Principles of Epidemiology in Public Health Practice. Third Edition. An Introduction to Applied Epidemiology and Biostatistics*. 2011: CDC.
24. Beaglehole R, Bonita R, Kjellstrom,T. *Basic Epidemiology*. 2006, Geneva: WHO. 174.
25. Colditz GA, Bonita R, Stampfer MJ, Willett WC, Rosner B, Speizer FE, Hennekens CH. *Cigarette smoking and risk of stroke in middle-aged women.* N Engl J Med, 1988. **318**(15): p. 937 – 941.

Chapter 9

Summary indicators of disease burden

Chapter objectives

In the previous chapters, we studied measures describing a single dimension of health or disease process, for example the outcome in the form of death or disease onset, evtl. number of diseases. On the other side, summary measures of population health refer to measures that combine information on mortality and non-fatal disease outcomes. The interest in the measures has been increasing during the last years, and calculation and publishing of particular measures has become a routine in many cases. Utilization of the measures in the area of public health has been discussed ever more, concurrently with the increasing interest in summary measures, starting with ethical consequences of social values inherent in the measures, technical and methodical issues related to the structure of various measures, concerns of distributive fairness and the use of summary measures as an input in decision-making on allocation of funds. [1]

Specialists from various areas of public health should be informed on the measures and be experienced in their interpretation. Since the use of such measures as indicators of various measures of activities and results of the public health processes has been continuously increasing, proper understanding of the mission provided by the measures is important for formulation of strategies on national, international but mainly on local and community level.

In general we recognize two types of summary health indicators: life-expectancy based group and health gaps- based group. Indicators of the first group extend the life expectancy idea with the fact that they don´t count only expected number of years to be lived up by a man but try to find out how many years will the man be able to live up in health (without disease or disablement). Indicators of health gaps are based on the idea of premature death and years lived with disease or disablement. [2] In this chapter, we will discuss in detail some summary indicators in the area of health gaps (PYLL and DALYs) and in the area of life expectancy (QAlYs and HLYs).

Potential Years of Life (PYLL)

Definition

Dictionary of Epidemiology [3] defines PYLL as a proportion of relative effect of various diseases and deaths on the society. It emphasizes the loss for the society as a result of deaths at young age or prematurely. Indicator PYLL applicable to particular death cause refers to the sum of years that people who died of it would have lived, had they lived up to particular age. We refer here to the premature death concept, i.e. death that occurred earlier than expected statistically. [4] To estimate the life expectancy, we use actuarial statistics we mentioned in preceding chapters. We still have to explain the age that will be taken as the limit, after which death will be considered normal. This decision should be made by the investigator himself, depending on the investigated problem. This is also the main problem of this indicator. [5]. Differences between two groups are not profound. In case of infectious diseases, chosen higher age limit results in rather small rise of PYLL contrary to much higher rise in case of cardiovascular diseases.

	1 - 70 years		**1 - 75 years**	
	PYLL (per 100,000)	**% (per 100)**	**PYLL (per 100,000)**	**% (per 100)**
Infectious diseases	655	3.0	797	2.8
Neoplasm	9008	41.7	11 933	41.2
Cardiovascular diseases	3 374	15.6	5 070	17.5
Cerebrovascular diseases	958	4.4	1 497	5.2
Respiratory diseases	780	3.6	1 196	4.1
Digestion track diseases	1 453	6.7	1 835	6.3
Other diseases	1 609	7.4	2 222	7.7
Violent deaths	3 812	17.6	4 385	15.2
All causes	21 649	100	28 935	100

Table No. 1 PYLL differences at various age limits. [5]

Some authors recommend using the life expectancy in order to restrict arbitrary chosen age limitation. [4] In such case, number of lost years is correlated to theoretical average number of years to be lived. The differences are not so big, as suggested by the table No. 1.

Indicator PYLL assigns more significance to deaths in younger age; the closer is the late´s age to the chosen limit, the smaller significance. For example, at chosen limit 70 years of age, death in age 65 contributes to PYLL with 5 but in case of death in age 55, contribution of potentially lost years refers to 15. Thus, the younger man dies, the more he contributes to PYLL value. While young people don´t die often, they significantly contribute to the lost life years´ indicator.

Calculation

There are a few PYLL calculation methods used in practice. Often, the age considered standard life length is determined upon the author´s decision. He can follow various objectives, thus decide on PYLL at his convenience. As well, various methods of determination of this parameter can be used. One of them is based on the constant value called *Normal Length of Life*, which expresses the age in which majority of people die. It is determined as a modus from the table count of the dead (e.g. 75 years for women and 70 years for men). Then, lost years by death refer to total years between the constant and actually reached age. Lost years are not counted at people who died later than the Normal Length of Life (NLL). Another method is based on the constant value called *Probable Length of Life*, or PLL It is the age *x* taken from the mortality table where it applies that table headcount of the live persons $lx » l0 / 2 = 50000$. In year 1999 it is the age 71 for men and 80 for women. Then, lost years by death refer to total years between the constant and actually reached age. Lost years are not counted at people who died later than PLL. The third method is based on the indicator *Potential Years of Life* or PYLL. In this case, lost years caused by death are determined as the average PYLL at the start and the end of the age interval when the person died. Such indicator is called also "life potential". [6] PYLL is usually expressed as recalculated per 100 thsd. citizens.

PYLL calculation procedure is stated in the Table No. 2. We decided to use standard life length 70 years for both genders, calculated from the 1st year of age.[30] The data source refers to the mortality statistics 2011 available from InFOstat.[31]

[30] It is possible to count in the age group < 1, with average death age 0.5. It is important to consider this group mainly if we calculate PYLL of diseases with high mortality rate up to 1 year of age.

[31] http://www.infostat.sk/slovakpopin/causes.htm

Age group	Age group mean	Remaining years a_i	Total deaths d_i	PYLL
1 – 4	3,0	67,0	0,0	0,0
5 – 9	7,5	62,5	0,0	0,0
10– 14	12,5	57,5	0,0	0,0
15– 19	17,5	52,5	0,0	0,0
20– 24	22,5	47,5	2,0	95,0
25– 29	27,5	42,5	4,0	170,0
30– 34	32,5	37,5	8,0	300,0
35– 39	37,5	32,5	21,0	682,5
40– 44	42,5	27,5	79,0	2172,5
45– 49	47,5	22,5	165,0	3712,5
50– 54	52,5	17,5	382,0	6685,0
55– 59	57,5	12,5	649,0	8112,5
60– 64	62,5	7,5	945,0	7087,5
65 – 69	67,5	2,5	1133,0	2832,5
In total (1-69)			3388,0	31850,0
Slovak population median (2011)				**4874348,5**
PYLL rate per 100 000				**653,4**

Table No. 2 Simple PYLL calculation per heart ischemia (ICD 10: I20-I25), Slovak citizens, 2011, data source INFOstat

We chose the group of diseases called heart ischemia (ICD 10: I20-I25). Calculation is easy and should be done in Excel or in {R} environment. At first, we have to calculate the average age of each group. We encounter problem with the very first group (1 – 4) since we have to correct the calculation, adding constant called Keyfitz approximation. [7] It refers to the fact that children face huge death risk immediately upon delivery, mainly in countries with high prenatal mortality rate but also in other countries with higher death risk during the first 28 days of life. This status can be compensated with correction of the mean value from original 2.5 to 3; however we have to admit some uncertainty. If we want to perform exact calculation, we have to use some of the approaches described by Keyfitz. Then we continue with calculation of the mean in particular age groups. Deducing the mean value of the groups from the standard life length, we will get potential years of life. Then we choose total death rate per all age groups and calculate PYLL for each of them, provided that we multiply the count of the dead by the remaining years of life ((70 –a_i)*d). Keeping patient, we can formulate it more exactly in the formula where d_i refers to headcount of the dead in the age group i, and a_i refers to years remaining till the target age 70. Calculation process described verbally above shall be expressed as multiplication of the dead´s count and the years remaining till the target age.

$$PYLL = \sum_{i=1}^{69} a_i d_i$$

We get the figure that shall be recalculated so as it expresses total years lost per 100 thsd. citizens (or other count). Thus, we shall divide the calculated value PYLL by headcount of citizens in particular year and in the age groups subject to our study, thus 1 – 69 years for both genders. The result shall be multiplied by the coefficient, namely by 100 000 in this case. The value reached refers to lost years of life per 100 000 citizens in particular age and cause.

PROS and CONS

PYLL advantage compared to life expectancy lays in simplicity of the calculation for certain groups of diseases, e.g. neoplasm or injuries. The indicator reflects mostly social values attributed to various death age categories. It expresses the significance of young people deaths because of lost potential future economic benefits for the society. [2] Another difference from the life expectancy represents possibility to sum up PYLL values and calculate total loss of years. Dependence of PYLL on the population size represents another feature. Thus, if we compare the life expectancy of two populations, its value will depend on the population size and the higher population count, the higher PYLL. The problem can be solved with calculation of the average PYLL value and then compare the values. [8] Common age category standardization can solve the problem. We can apply direct standardization to standard population and the process whose principle was described in the chapter dealing with standardization.

Standardization

PYLL value is only seldom used in the real life; usually it serves for comparison of two populations. In case of populations of various age structure, we will need standardization. As we said in the chapter dealing with standardization, we can do it using a new standard population introduced for WHO. [9]

To describe the PYLL standardization method, we will use the data applicable to heart ischemia (ICD 10: I20-I25), Slovak citizens, 2011, and perform calculation in the format Excel. To perform PYLL standardization, we have to find out death count

Age group	Average age	Remaining years	Total deaths	Gross $PYLL_i$	Slovak population median	Standard population	Expected count in standard population u_{wi}	Age standardized $PYLL_{wi}$
		a_i	d_i	$d_i * a_i$	n_i	w_i	d_i/n_i*w_i	$d_{wi} * a_i$
1 – 4	3,0	67,0	0,0	0,0	227073	7033	0,00	0,00
5 – 9	7,5	62,5	0,0	0,0	263932,5	8687	0,00	0,00
10 – 14	12,5	57,5	0,0	0,0	281904,5	8597	0,00	0,00
15 – 19	17,5	52,5	0,0	0,0	340498,5	8474	0,00	0,00
20 – 24	22,5	47,5	2,0	95,0	399741,5	8222	0,04	1,95
25 – 29	27,5	42,5	4,0	170,0	437056,5	7928	0,07	3,08
30 – 34	32,5	37,5	8,0	300,0	459620	7605	0,13	4,96
35 – 39	37,5	32,5	21,0	682,5	429985	7145	0,35	11,34
40 – 44	42,5	27,5	79,0	2172,5	358124,5	659	0,15	4,00
45 – 49	47,5	22,5	165,0	3712,5	373723	6038	2,67	59,98
50 – 54	52,5	17,5	382,0	6685,0	383588	5371	5,35	93,60
55 – 59	57,5	12,5	649,0	8112,5	386354	4547	7,64	95,48
60 – 64	62,5	7,5	945,0	7087,5	313074	3723	11,24	84,28
65 – 69	67,5	2,5	1133,0	2832,5	219673,5	2955	15,24	38,10
In total (1-69)			3388,0	**31850,0**				**396,79**
Slovak population median (2011)				**4874348,5**	Standard population in total			**86984**
PYLL rate per 100 000				**653,4**	Age standardized PYLL Per 100 000			**456,16**

Table No. 3 Direct PYLL standardization resulting from heart ischemia (ICD 10: I20-I25) in 2011.

expected in standard population (w_i), where people die according to our population mortality rate. Such expected count (d_{wi}) shall be multiplied with the remaining life years (a_i). PYLL results from the sum of the multiplication, which shall be further divided by total headcount of persons in standard population and multiplied by index number. The result is stated in the table No. 3.

Disability adjusted life years (DAlYs)

Definition

Speaking of PYLL, we considered disease burden expressed by the number of premature deaths. There are situations, however, when a person suffering diseases survives for a long time (e.g. diabetes, blindness, deafness, immobility after injury) and is restricted by it for a long time. How could the effect of such conditions on the population health be expressed generally? In the 90s of the 20th century, indicator was created within the study called Global Burden of Disease, Injuries and Risk Factors (GBD), aimed at evaluating global burden of disease – disability-adjusted life years, (DALYs). DALYs represents the population health indicator that combines lost life years resulting from premature death and life years lived by a person with worsened health condition. [10, 11] The indicator has been used for example by WHO in its publications dedicated to global population disease burden. [12] Since publishing the first results of GBD during1996/7, the document was a few times updated. [13] Prior to starting description of the DALYs calculation, we should define the term disability. According to Murray et al. [13] it is any short- or long-term loss of full health other than death (pain, immobility, lost cognitive functions, etc.). Reader could ask why it is necessary to express the disease burden with a single summary indicator. Authors of the indicator [13] enlisted at least three reasons why summary health indicators are necessary:

1. They can serve for comparison of population or community health in particular time and space.
2. Their application allows for making a broader image of the extent of attention to be paid to particular diseases. This would allow competent authorities to decide on and address major problems, and monitor whether these are resolved or more complicated based on the interventions. Mental diseases represent an example of such broader image; since they don´t cause many fatalities but if we look at them via DALYs, they become a major issue (thus, if we examined the disease burden only from mortality rate, we wouldn´t consider it a problem) (Table No. 4)

Mortality	DALYs
Heart ischemia	Lower respiration track infections
Cerebrovascular diseases	Diarrhea and the like
Lower respiration track infections	Depression
Chronic obstruction lung disease	Heart ischemia
Diarrhea and the like	HIV/AIDS
HIV/AIDS	Cerebrovascular diseases
Tuberculosis	Premature birth and low birth weight
Lung and bronchial tumors	Asphyxia and trauma at delivery
Traffic accidents	Traffic accidents
Premature birth and low birth weight	Neonatory infections

Table No. 4 First 10 mortality and DALYs causes – all ages, 2004. Source: [12]

3. Their application can provide us with information on the routine data collection weak points. If we find out that we are unable to calculate the indicator but it can be done in other country, it is a proof of input data collection system deficiencies in our country.

Calculation and interpretation

DALYs calculation has slightly changed with increasing number of GBD updates. We will describe the method applied by WHO to the update in 2004 [14], however GBD 2010 update used slightly different method. [13] We know that DALYs is an indicator expressing population disease burden in the form of premature death and non-fatal consequences with a single digit. Thus, we need two components to calculate DALYs. The first one expresses premature mortality and another one refers to worsened health condition as a result of particular disease. DALYs is calculated per particular psychical or physical disablement caused by disease or injury in particular age group. [15]. DALYs calculation formula [14]:

$$DALY = YLL + YLD,$$

where YLL represents life years lost as a result of premature death and YLD years lived with disability, years lost due to disability.

Variables YLL and YLD should be inserted in the basic formula. Look at first at YLL – life years lost as a result of premature death caused by particular disease. It is calculated according to the following formula:

$$YLL = N * L,$$

where N refers to all deaths of particular disease at certain age (age group) and L is standard life expectancy at the death age/ age group, expressed in years.

As we see, value characterizing the life expectancy of a person is required to calculate this component. There are questions frequently asked, which life expectancy should be used by the investigator. All GBD but the latest one dated 2010 used the life expectancy of Japanese women at the delivery (the highest world value referring to 82.5 years). Men´s life expectancy was reduced to 80 years, reflecting the biological differences of both genders. [11] Other organizations (e.g. CDC) apply the age 75 years. Of course, application of different values results in different YLL (and thereby also DALYs) values, which should be taken in account when choosing the methodics (for example, one should consider possibility to compare our and other authors´ values) [16].

Age category	Standard life expectancy West Level 26	
	Men	Women
0	79,94	82,43
1-4	77,77	80,28
5-9	72,89	75,47
10-14	67,91	70,51
15-19	62,93	65,55
20-24	57,95	60,63
25-29	52,99	55,72
30-34	48,04	50,83
35-39	43,10	45,96
40-44	38,20	41,13
45-49	33,38	36,36
50-54	28,66	31,68
55-59	24,07	27,10
60-64	19,65	22,64
65-69	15,54	18,32
70-74	11,87	14,24
75-79	8,81	10,59
80-84	6,34	7,56
85+	3,54	4,25

Table No. 5 Standard life expectancy used by WHO to calculate YLL (interpolation from detail tables)

Table No. 5 contains the mentioned standard life expectancy used by WHO for calculation of DALYs in all countries.[32] Reader could wonder why the life expectancy of particular country isn´t applied. It´s because the DALYs concept uses the principle of equality. It means that all studies should use the same "ideal" life expectancy regardless race, social-economic status or occupation (age and gender are the only factors considered).

After YLL calculation, we can move to calculation of years of life disability (YLD). The figure defines the burden laid on the population by disease in the form of illness rate. YLD is calculated from the formula:

---

[32] Data files can be downloaded from: http://www.who.int/healthinfo/global_burden_disease/tools_national/en/index.html

$$YLD = I * DW * L,$$

where I is the number of new (incidental) cases of disease, DW is the disablement weight and L is an average illness duration.

Alternatively, YLD can be calculated using prevalence according to the following formula [13, 17]:

$$YLD = P * DW,$$

where P refers to the number of the ill in the population and DW is the disablement weight.

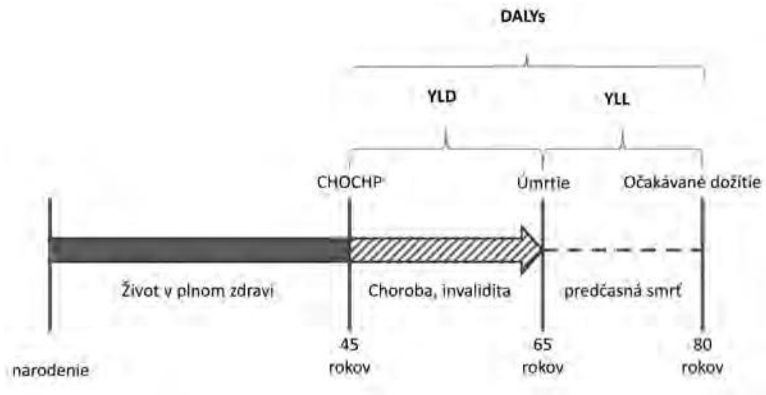
When discussing YLD, it is necessary to define the input data, referring to the number of new disease cases in particular age-gender group per particular period; the disablement weight and average illness duration (until healing or death). The source of new cases incidence are probably identifiable (we discussed few of them in the previous chapters) but such data could be hard to obtain in case of chronic diseases. Questions could appear regarding the last two data. Disablement weight is a factor expressing disease significance with the scale 0 (full health, no health loss) and 1 (death, total health loss) [18]. Disablement weight is subjective information, definitely, created on basis of the experts´ or scientists´ panel opinion [16, 19]. It should be obtained from official sources, for example from WHO[2]. GBD 2010 used other disablement weights. [19] Table No. 6 contains WHO-used disablement weights for particular diseases.

Disease / health disorder	**Disablement weight (untreated forms)**	**Disablement weight (treated forms)**
Terminal cancer stage	0,809	0,809
Diabetic foot	0,137	0,129
Blindness (glaucoma effect)	0,600	0,600
Angina pectoris	0,227	0,095
Chronic obstructive pulmonary disease (COPD)	0,428	0,388
Asthma	0,099	0,059

Table No. 6 Examples of disablement weights used in the Global Burden of Diseases 2004

Average disease duration is also a speculative data to be drawn from the literature.

Let´s make an example of DALYs calculation per 1 person. We want to express the value of DALY caused by untreated COPD to a man who fell ill with COPD in 45th year of age and died as 65 years old of the same disease. Look at Picture No. 1 illustrating the case in question.



Picture No. 1 Schematic illustration of DALYs

The man fell ill in the age of 45; before he had been completely healthy. He died as 65 years old. Since we know that he should have lived until 80, we will find out through simple deduction that the man lost 15 years of life as a result of premature death. Thus, we can apply already known formula thereto:

$$YLL = N * L = 1 * 15 = 15.$$

Value YLL refers to15 lost years resulting from premature death. Now we can calculate YLD. Since the case involves the only person, number of new cases refers to 1, untreated COPD disablement weight is found in the table No.5 and refers to 0.428. In this case, disease duration is easy since we see from the assignment and scheme that he had suffered COPD for 20 years. We insert the data in the formula:

$$YLD = 1 * DW * L = 1 * 0.428 * 20 = 8.56.$$

Thus, the man spent 8.56 years of his life in the health disablement.

Summing up YLL and YLD, we will get DALYs of value 23.56. Therefore, the man suffered 23.56 DALYs as a result of untreated COPD. This indicator should be interpreted similar to other indicators expressing adverse effects (death, disease), thus DALYs should be prevented from. The smaller is DALYs the better is the situation. Intervention cost analysis refers to one of the DALYs applications. The same disablement weight table states also disablement weight of treated COPD (0.388). We could ask how many DALYs would treat COPD cause to the same man; or how many DALYs would have the man saved if he had been treated. In such case, YLL remains unchanged – 15 years. YLD will be calculated with insertion of treated COPD disablement value in the formula:

$$YLD = 1 * DW * L = 1 * 0.388 * 20 = 7.76.$$

Summing up YLL and YLD we will get DALYs of value 22.76 for our man in case of treated COPD. Thus, had the man been treated, he would have saved 23.56 – 22.76 = 0.80 DALYs. If we knew the treatment price, we could calculate required cost of one DALY saving.

Similarly, we could calculate DALYs of the whole population for particular age-gender groups and diseases. If we want to calculate DALYs for the whole population, for COPD in particular year, we will need to know the count of the persons who died of COPD and the count of new COPD cases in the monitored age-gender category, the disablement weight and average disease duration, usually drawn from literature. The resulting DALYs are usually presented similar to PYLL per certain headcount (e.g. per 100 000 citizens).

We see that many indicators enter the DALYs calculation process, resulting from estimations or surveys, thus with bias. Therefore we shouldn´t be satisfied with the only DALY value. The problem can be solved with method Monte Carlo that will be discussed in detail in the chapter dealing with mathematical models.

Above presented DALYs calculation has been simplified and doesn´t take in account age weight and time discount. (These were included in all GBD except GBD 2010). Briefly, we could state that age weighting means in general that certain societal preferences were considered in the DALYs calculation, which means that death and health disablement of a young person will be significantly highlighted during calculation (10 – 55 years) compared to children and elderly. Time discounting means that a year spent in better health has currently higher value than the same in the future. Introduction of these preconditions in the formula makes it much more complicated but we are not going to discuss it since it´s not been used frequently because of huge criticism. People interested in this subject matter can study the Murray´s article dated 1994 [11] and the reasons of skipping these preconditions in the article dated 2012. [20]

PROS vs CONS

DALYs pros and cons are clearly understandable for a reader from the above text. The indicator PROS refer mainly to 3 arguments of its authors, stated at the beginning. The indicator CONS represents certain complicatedness of calculations and the need for making presumptions (disablement weight and average disease duration) or application of standard life expectancy. If we look at the disablement weight, there are authors saying that all disablement weights shouldn´t be applied to all persons; arguing that being blind in an undeveloped country is worse than in a developed country. [21] DALYs in countries of shorter life expectancy is also artificially increased by application of standard life expectancy. [22] Thus, it is necessary to specify which presumptions were inserted in the calculation when presenting DALYs.

Since the indicator DALYs has been ever more used during description and comparison of population health condition, mainly when speaking of disease burden, public health experts should be aware of the indicator interpretation value. We don´t think it is necessary to know the calculation details but knowing DALYs pros and cons will definitely contribute to development of national health policies. The calculations will be for sure made by demography and health statistics experts.

Quality adjusted life years (QAlYs)

Definition

Pressure on the finances in healthcare industry has been increasing together with fast growing possibilities of diagnostics and treatment of diseases. It is complicated to decide whether a new and mostly costly treatment should be applied but such decisions must be made. Thus, the way has been sought how to express particular treatment benefits. One of the possibilities refers to application of indicator called "quality adjusted life years" or QAlYs.The indicator was introduced in the 2nd half of the 20th century as a cost analysis efficiency indicator. This method should help people competent to decide on allocation of limited funds to various projects and health programs [23]. Currently, QAlYs is used for comparison of various treatments´ benefits or comparison of treatment vs non-treatment benefits. QAlY is an indicator taking in account both life dimensions: quantity and quality. QAlY expresses years of an individual spent in full life quality (1 year of life spent in full quality = 1QAlY, 1 year of life spent in quality of weight 0.25 = 0,25 QAlY). [24]

Calculation

Calculation of QAlYs is rather simple. We will explain it on the example of patient suffering chronic disease and two treatment forms [25]. If we should decide on the form of treatment the patient will get we have to answer a couple of questions. Would both treatments prolong the patient´s life: If yes, by how much? What are adverse effects of the treatments? (How will they affect the patient´s life quality? Last but not least: How much do the treatments cost? Basic formula for calculation of QAlYs per an individual:

$$QALY = N * Q,$$

where N refers to total years lived in particular health condition and Q refers to life quality associated with the health condition.

Similar to DALYs, quality weight is expressed with numbers 0 – 1.

In case of QAlYs, 0 represents the worst life quality (death) and 1 represents highest health quality. Moreover, the quality weight can be of negative value as well, thus even worse than death (e.g. staying in bed with extreme pain). Similar to DALYs disablement weights, quality weights are obtained from specialized survey conducted with the population or a group of patients. Most common methods of data acquisition refer to *standard gamble* technique (respondents have to choose between maintaining current status for 10 years or try a treatment that could completely heal them or cause death [26]), *time trade-off* (respondent should choose between standard life length in worsened health condition or reduced life length in full health [26]) or *visual range* (where he subjectively describes his actual life quality). [22, 24] Questionnaire *EQ-5D* represents another possibility to obtain quality weight for our calculations [24].

Come back to our example. If our patient is treated by therapy A, he will live one more year and his life quality will refer to 0.4. If he is treated by therapy B, he will live 1 year and 3 months (1.25 years) with life quality 0.6.

We should compare therapy A and B, finding out how many QAlYs the patient gets from both therapies:

Therapy A: $QALYs = 1 * 0{,}4 = 0{,}4$
Therapy B: $QALYs = 1.25 * 0.6 = 0.75$

PROS vs CONS

You see that therapy B will bring more QAlYs to the patient – the difference refers to 0.35 QAlYs. If we can calculate QAlYs, we can further consider the cost effectiveness and ask how much would therapy cost to get one single QAlY? Let´s say that therapy A costs 3 000 € and therapy B costs 10 000 €. The cost difference (7 000 €) shall be divided by QAlYs difference (0.35) and the result refers to the therapy price of 1 QAlY. It is 20 000€. We should ask when the price of 1 QAlY is too high. Despite of evaluating each case separately, there are recommendations establishing the limit beyond which the therapy price is too high. As an example, we state the British NICE (national Institute for Health and Care Excellence), to which therapy exceeding cost GBP 20 000 – 30 000 per 1 QAlY is considered cost ineffective. [25]

While quantity in QAlYs context is relatively easy measureable (we can measure life duration/ length), life quality measurement is rather demanding since it has various meanings. Moreover, QAlYs concept doesn´t take in account quality of other people´s life (family relatives) that care about the ill. QAlYs concept presumes that life quality prevails life quantity. While many people accept such presumption, other patients would accept therapy side effects if it was able to prolong their life. There are opinions that one of QAlYs advantages is that it doesn´t handicap people on the age basis (therapy of 75-years old man with 5 year life expectancy has the same priority as a therapy of 40-years old man in terminal disease stage also with 5 year life expectancy and the same life quality). Other opinions state that just this feature is a negative since 40-years old man should have higher priority than the 80-years old one, because the latter already had opportunity to live longer. While QAlYs are frequently applied, some authors criticize them as subjective and discuss the ethics of using them; stating that it allows for potential decision on who is and who isn´t entitled to the therapy. [24]

Healthy lived years (HLYs) or disablement-free life expectancy (DFLE)

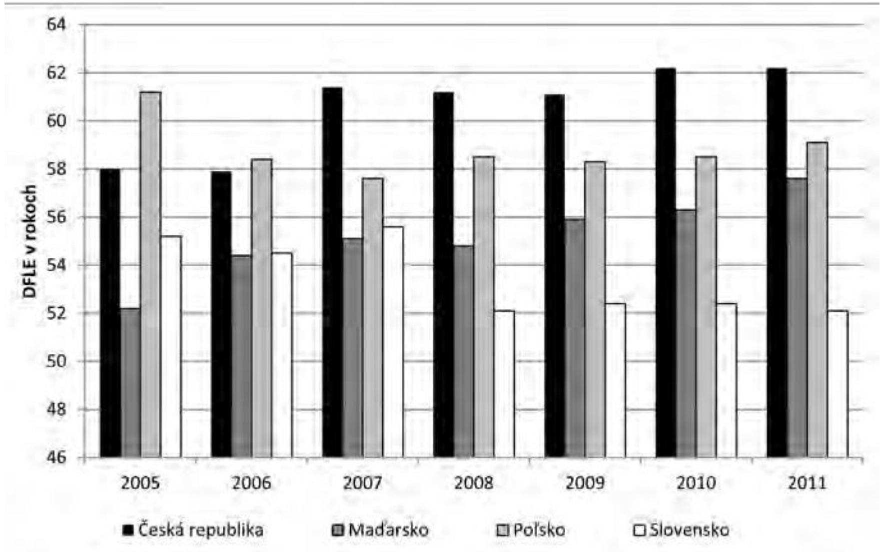
Definition

Last summary indicator of population health we are going to discuss is Disability-Free Life Expectancy (DFLE) or Healthy Life Years (HLYs). In the chapter dealing with life expectancy, we said we can calculate how many years should a child born in this year live up to, if the population mortality trends taken as the basis of the calculation, remained unchanged. Thus, we are able to estimate the life expectancy but we should ask how many years the child will be healthy (i.e. not restricted by any disease, injury or other health disorder). DFLE serves to this purpose. To define it correctly, DFLE (HLYs) is an indicator measuring number of years during which the person of certain age remains healthy (without disablement). The indicator takes in account the fact that a person usually doesn´t live the whole life in full health. [27] DFLE is usually calculated by Eurostat[33] (Picture No. 2).

Calculation

The Sullivan´s method dated 1971 is used for calculation of the indicator. [28] The method is based on the age-specific disablement prevalence in the population and mortality rate/ data. If we want to calculate DFLE, we need the death tables similar to the life expectancy calculation and information on the age-specific disablement prevalence in the population. While the first component is easily accessible, the age-specific disablement prevalence in the population is hard to obtain and can be provided by the population survey. According to Mésáros, EU-Silc (*Statistics on Income and Living Conditions*) survey is available in Slovakia, performed on annual basis since

[33] http://epp.Eurostat.ec.europa.eu/portal/page/portal/health/public_health/data_public_health/database



Picture No. 2 "Life expectancy in health upon birth in four EU countries between 2005 – 2011.
Source: Eurostat

2004. He states three questions from the survey that could serve for purpose of DFLE calculation. We will show the calculation using data similar to Eurostat.[34] The question is: *„Did you have to restrict your activities for health reasons during the last 6 months at least?"* with optional answers: *„I didn´t"*, *„I had to, partly"*, *„I didn´t have to"*. [29] We will make calculation of DFLE on the example of Slovak men in 2011. All data required for the calculation are available on website EurOhex (*European Health Expectancies*), containing DB of indicators and data necessary for calculation of life expectancy, as well as HLYs for 27 EU countries. [35] Detail calculation description and its mathematical principles are described by Mészáros and the instruction manual dated in 2007. [29, 30]

At first we will get necessary data: detail death tables and disablement prevalence. Both components can be found in the web DB. In section „Life tables" we choose „national data" and adjust our selection so as we get complex detail death tables for all men´s ages in Slovakia in 2011. We can see that life expectancy of men born in 2011 refers to 72.3 years (Picture No. 3). Then we export the tables in Excel.

---

[34] http://epp.Eurostat.ec.europa.eu/cache/ITY_SDDS/en/hlth_hlye_esms.htm
[35] http://www.eurohex.eu/IS/

Life table

Life table in SLOVAKIA by year for All ages and Men

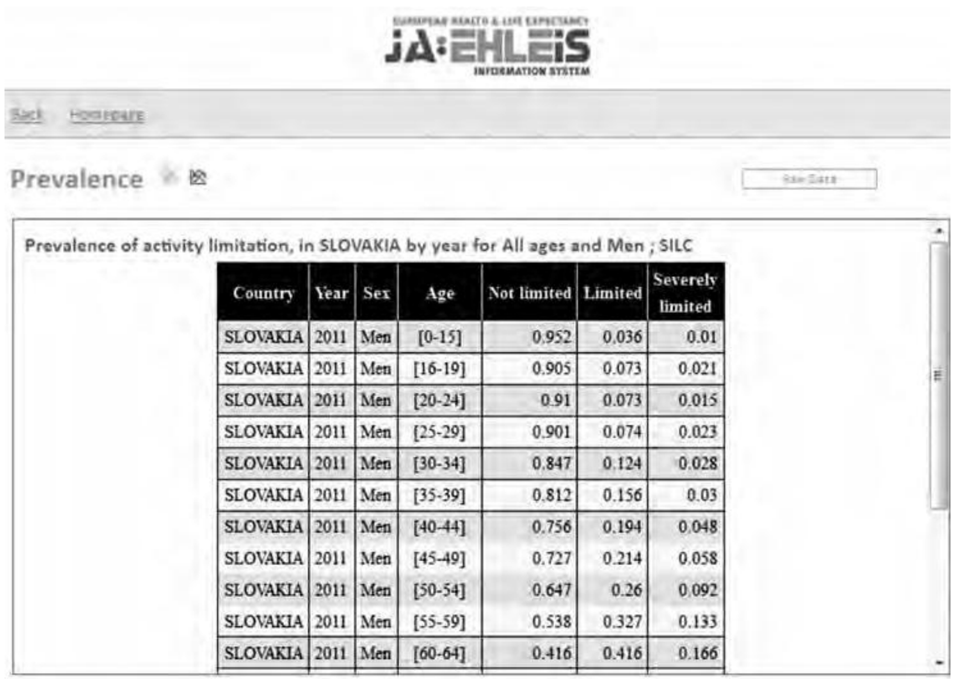
Country	Year	Sex	Age	mx	qx	lx	dx	Lx	Tx	ex
SLOVAKIA	2011	Men	0	0.00571371	0.00569743	100000	569.743	99544.204	7230338.755	72.3
SLOVAKIA	2011	Men	1	0.00043295	0.00043285	99430.256	43.039	99408.736	7130794.55	71.7
SLOVAKIA	2011	Men	2	0.00016628	0.00016627	99387.216	16.525	99378.954	7031385.814	70.8
SLOVAKIA	2011	Men	3	0.00027953	0.0002795	99370.691	27.774	99356.804	6932006.86	69.8
SLOVAKIA	2011	Men	4	0.0002872	0.00028716	99342.917	28.527	99328.653	6832650.055	68.8
SLOVAKIA	2011	Men	5	0.00028652	0.00028648	99314.389	28.451	99300.163	6733321.402	67.8
SLOVAKIA	2011	Men	6	0.00010749	0.00010748	99285.937	10.671	99280.601	6634021.238	66.8
SLOVAKIA	2011	Men	7	0.00029462	0.00029458	99275.266	29.244	99260.643	6534740.636	65.8
SLOVAKIA	2011	Men	8	0.00026612	0.00026609	99246.021	26.408	99232.816	6435479.992	64.8
SLOVAKIA	2011	Men	9	0.00022761	0.00022759	99219.612	22.581	99208.321	6336247.176	63.9
SLOVAKIA	2011	Men	10	0.00014574	0.00014573	99197.03	14.456	99189.802	6237038.854	62.9
SLOVAKIA	2011	Men	11	0.00010576	0.00010576	99182.574	10.489	99177.329	6137849.051	61.9

Picture No. 3 Output from DB EurOhex – death tables for men in Slovakia in 2011

Furthermore, we need information on the disablement prevalence. In the DB, choose section „Health data“, then item „national data“, and the last option – EU-Silc survey results and within it, the option “activity restriction”. We should set forth the requirements so as we can obtain information on the disablement prevalence for men in Slovakia in 2011 (Picture No. 4). Then again export the data in Excel.

Now we can see that the prevalence data structure (0–15, 16–19...85+) doesn´t correspond to the data structure in the death / life tables (0,1,2,3,4...85+). Never mind, there are two versions of the problem solving. We can use either detail life tables and presume that the prevalence is equal in every age included in particular age interval (thus serious disablement prevalence – 0,01, observed at age group 0 – 15 will be similar to the age group 0,1,2,3,4....15), or we can apply the method of abridged life tables and adjust them to the prevalence data format [30]. We have chosen the 1st method.

On the picture No. 5, we can see a complete calculation of life expectancy in health of men in Slovakia in 2011. In fact, it is combination of data chosen from DB. To simplify the calculation description, we marked rows and columns. Life table refers to the 1st table part from the column [2] up to [8] (only a part of the detail table is displayed because of its size). We inserted disabled people prevalence in the column [9] (partial and major disablement) at all ages.



Country	Year	Sex	Age	Not limited	Limited	Severely limited
SLOVAKIA	2011	Men	[0-15]	0.952	0.036	0.01
SLOVAKIA	2011	Men	[16-19]	0.905	0.073	0.021
SLOVAKIA	2011	Men	[20-24]	0.91	0.073	0.015
SLOVAKIA	2011	Men	[25-29]	0.901	0.074	0.023
SLOVAKIA	2011	Men	[30-34]	0.847	0.124	0.028
SLOVAKIA	2011	Men	[35-39]	0.812	0.156	0.03
SLOVAKIA	2011	Men	[40-44]	0.756	0.194	0.048
SLOVAKIA	2011	Men	[45-49]	0.727	0.214	0.058
SLOVAKIA	2011	Men	[50-54]	0.647	0.26	0.092
SLOVAKIA	2011	Men	[55-59]	0.538	0.327	0.133
SLOVAKIA	2011	Men	[60-64]	0.416	0.416	0.166

Picture No. 4 Output of DB EurOhex on prevalence of disablement among Slovak men in 2011

As we said, we presume that prevalence available only for age intervals and included in particular interval is equal in every age. We have to calculate headcount of persons at particular age living without disablement in our model population (column [10]). This can be calculated through multiplication of headcount of people living at the end of the age with proportion of people living disablement-free at the same age, i.e.[6] * (1 – [9]). Then we repeat the calculation for the whole life table (up to age 85+). The column [11] – the years lived by people at particular age without disability, shall be calculated when summing up all values in the column [10] starting from the applicable age up to the highest age in the life table (85+). It can be simplified as follows: [10A]+ [11B] and further up to the table end. Then we divide total years lived without disablement [11] at every age with respective headcount of persons in the model population [4], and we get DFLE [12]. Now we can see that the Slovak men had life expectancy in health upon birth corresponding to 52.2 years, while 65-years old man faces 3.5 years of healthy life more. The column [13] states percentage of healthy life years from the whole life expectancy. We can calculate it when dividing column [12] by column [8]*100.

PROS vs CONS

Advantages of the indicator are similar to life expectancy. It is easy to

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
	Vek	Úmrtnosť	qx	lx	Lx	Tx	LE	Proporcia ľudí žijúcich bez invalidity vo veku x	Osoby žijúce bez invalidity	Celkové roky života prežité bez invalidtiy od veku x	DFLE	% rokov bez invalidity z LE
[A]	0	0,0057	0,0057	100000,00	99544,20	7230338,76	72,3	0,952790	94844,69	5216729,91	52,2	72,2
[B]	1	0,0004	0,0004	99430,26	99408,74	7130794,55	71,7	0,952790	94715,62	5121885,22	51,5	71,8
[C]	2	0,0002	0,0002	99387,22	99378,95	7031385,81	70,7	0,952790	94687,24	5027169,61	50,6	71,5
[D]	3	0,0003	0,0003	99370,69	99356,80	6932006,86	69,8	0,952790	94666,13	4932482,37	49,6	71,2
[E]	4	0,0003	0,0003	99342,92	99328,65	6832650,06	68,8	0,952790	94639,31	4837816,24	48,7	70,8
...	...	...	...	...	...	...	...	...	...	...	...	...
...	61	0,0205	0,0203	80686,40	79867,06	1378317,31	17,1	0,416843	33292,04	383782,94	4,8	27,8
...	62	0,0227	0,0225	79047,72	78158,96	1298450,25	16,4	0,416843	32580,03	350490,90	4,4	27,0
...	63	0,0241	0,0238	77270,20	76350,97	1220291,29	15,8	0,416843	31826,38	317910,87	4,1	26,1
...	64	0,0247	0,0244	75431,74	74511,12	1143940,32	15,2	0,416843	31059,45	286084,49	3,8	25,0
...	65	0,0273	0,0269	73590,50	72600,70	1069429,20	14,5	0,365543	26538,69	255025,04	3,5	23,8
...	...	...	...	...	...	...	...	...	...	...	...	...
...	81	0,1055	0,1002	31378,57	29806,04	200471,80	6,4	0,162975	4857,65	19283,31	0,6	9,6
...	82	0,1192	0,1125	28233,51	26645,91	170665,76	6,0	0,162975	4342,62	14425,67	0,5	8,5
...	83	0,1299	0,1220	25058,31	23529,52	144019,86	5,7	0,162975	3834,73	10083,05	0,4	7,0
...	84	0,1356	0,1270	22000,72	20604,08	120490,34	5,5	0,162975	3357,95	6248,32	0,3	5,2
...	85	0,1923	0,1754	19207,43	99886,26	99886,26	5,2	0,028937	2890,36	2890,36	0,2	2,9

Picture No. 5 Example výpočtu očakávanej dĺžky života v zdraví pre slovenských mužov v roku

interpret and understand. Along with complicated calculation, its disadvantage refers to the use of information from population surveys that can be of different quality. Accordingly, data from various surveys can be used (not only EÚ-Silc), and calculated indicators can be incomparable among the countries. For such reasons, WHO and EU strive for standardization of the population surveys. EHIS (European Health Interview Survey) is a good example thereof; the countries monitor subjective health condition on regular manner, applying the same methodology. Details of EHIS study conducted in Slovakia are contained on the website of the Slovak Statistical Office.[36] Detail description of the methodology and its latest updates are available on Eurostat website.[37]

Summary

In this chapter, we discussed summary indicators describing the disease burden on the population in the form of illness and mortality rate. We started with the discussion of PYLL, expressing the disease burden laid on the young population mainly in the form of premature death. Furthermore, we discussed indicators DalYs and QAlYs, which can be used also in the area of pharmaceutical economy and evaluation of health support program efficiency. The chapter was concluded with the indicator HLYs (DFLE), representing an extension of the life expectancy concept. We enlisted pros and cons for each indicator that should be taken in account to provide their rational and correct interpretation.

References

1. Murray, C. J. L., Salomon, J. A., Mathers, C. D., Lopez, A. D. *Summary Measures of Population Health Concepts, Ethics, Measurement and Applications*. 2002, Geneva: World Health Organization
2. Young, T. K. *Population Health: Concepts and Methods*. 2nd ed. 2004: Oxford University Press, USa.
3. Last, J. M. *Dictionary of Epidemiology*. 1999, Bratislava: USAID.
4. Manton, K. G., Akushevich, I., Kravchenko, J. *Cancer Mortality and Morbidity Patterns in the U.S. Population: An Interdisciplinary Approach*. Statistics for Biology and Health. 2009: Springer Science Business Media, LLC.
5. Meslé, F. *Medical Causes of Death*, in *Demography: Analysis and Synthesis, Four Volume Set, Volume 1-4: A Treatise in Population*, G. Caselli, G. Wunsch, and J. Vallin, Editors. 2005, Academic Press. p. 2976 pages

[36] http://portal.statistics.sk/showdoc.do?docid=28834

[37] http://epp.Eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:European_health_interview_survey_%28EHIS%29

6. Mészáros, J. *Calculation of Life Tables and Lost Years by Death*. 2000. 14.
7. Keyfitz N. *Introduction to the Mathematics of Populations*. 1968: Addison-Wesley Educational Publishers Inc
8. Webb, P., Bain, C. *Essential Epidemiology: An Introduction for Students and Health Professionals*. 2010, Cambridge: University Press. 458.
9. ahmad O et al. *Age standardization of rates: a new WHO standard*. 2001 WHO EIP/GPE/EBD, 14.
10. Lopez, D. A. et al. *Global Burden of Disease and Risk Factors*, 2006, Oxford University Press, World Bank: New York, Washington, DC. p. 475.
11. Murray, C. J. *Quantifying the burden of disease: the technical basis for disability-adjusted life years*. Bulletin of the World Health Organization, 1994. **72**(3): p. 429 – 445.
12. WHO. *The Global Burden of Disease: 2004 Update*, 2008.
13. Murray, C. J. L. et al. *Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010 (vol 380, pg 2197, 2012)*. Lancet, 2013. **381**(9867): p. 628 – 628.
14. WHO. *National Burden of Disease Studies: A practical guide*, 2001: Geneva.
15. Fox-Rushby, J. A., Hanson, K. *Calculating and presenting disability adjusted life years (DALYs) in cost-effectiveness analysis.* Health Policy and Planning, 2001. **16**(3): p. 326 – 331.
16. Steenland, K., Armstrong, B. *An overview of methods for calculating the burden of disease due to specific risk factors.* Epidemiology, 2006. **17**(5): p. 512 – 519.
17. WHO. *Metrics: Disability-Adjusted Life Year (DALY)*. 2013 [cited 2013 19.7.]; Available from: http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/.
18. WHO. *Disability weights, discounting and age weighting of DALYs*. 2013 [cited 2013 19.7.]; Avail-able from: http://www.who.int/healthinfo/global_burden_disease/daly_disability_weight/ en/index.html.
19. Salomon, J. A. et al. *Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010.* Lancet, 2013. **380**(9859): p. 2129 – 2143.
20. Murray, C. J. L. et al. *GBD 2010: design, definitions, and metrics.* Lancet, 2012. **380**(9859): p. 2063 – 2066.
21. Robberstad, B. *QALYs vs DALYs vs LYs gained: What are the differences, and what difference do they make for health care priority setting?* Norsk Epidemiologi, 2005. **15**(2): p. 183 – 191.
22. Sassi, F. *Calculating QALYs, comparing QALY and DALY calculations.* Health Policy and Planning, 2006. **21**(5): p. 402 – 408.
23. Weinstein, M. C., Torrance, G., McGuire, A. *QALYs: The Basics.* Value in Health, 2009. **12**: p. S5 – S9.
24. Kirkdale, R. et al. *The cost of a QALY.* QJM, 2010. **103**(9): p. 715 – 720.
25. NICE. *Measuring effectiveness and cost effectiveness: the QALY*. 2010 20.4.2010 [cited 2013 21.7.]; Available from: http://www.nice.org.uk/newsroom/features/measuringeffectivenessandcostef-fectivenesstheqaly.jsp.
26. Porta, M. *A dictionary of Epidemiology. Fifth Edition*. 2008, Oxford: Oxford University Press.
27. European Committee. *Healthy Life Years (HLY)*. 2013 22.7. [cited 2013 22.7.]; Available from: http://ec.europa.eu/health/indicators/healthy_life_years/hly_en.htm.
28. Sullivan, D. *A single index of mortality and morbidity.* HSMHA Health Rep., 1971. **86**(4): p. 347 – 354.
29. Mészáros, J. *Calculation of Health Life Median (methodical material)*, 2008, InFOstat: Bratislava.
30. Robine, M. J. et al. *Health Expectancy Calculation by the Sullivan Method: A Practical Guide. 3rd Edition*, 2007.

Chapter 10

Statistics of healthcare services

Chapter objectives

The state healthcare cost has increased since 2004 from Eur 3.5 bn to Eur 5.9 bn in 2010. It was reflected in the population health condition in the form of increased life expectancy at birth from 74.1 to 75.2 years. Thus, we paid by Eur 2.4 bn more for prolonged life expectancy by 1.1 year during 10 years, among other costs. Is it much or not? What did we spend the money for? These questions can be answered only when knowing the population demography in its dynamics, population health condition, healthcare services structure and activities, and respective sources. This chapter is dedicated to healthcare services statistics. It is an extensive area and therefore we are not able to discuss all its components and levels here. Healthcare data represent a major part of the views on the public health; describing available capacities required for healthcare provision but hiding also potential obstacles. They reflect quantity but enable considerations of quality of the provided healthcare. Last but not least, the data point out to imbalance in the approach to the services and provide the grounds for sources sustainability considerations.

We will keep to the indicators used by Eurostat and the national center of healthcare information. Accordingly, we will put aside a sub-chapter dedicated to the healthcare facility chain and activities. Human resources are discussed in other chapter. Economic indicators are defined within the *National Healthcare Accounts*, representing a part of the *System of National Accounts* [1] in a separate sub-chapter.

Similar to the previous chapters, we will present the data from the Slovak sources and attach them with European sources as required. Having studied the chapter, a reader should become familiar with the healthcare service recordkeeping system in Slovakia and EU and be able to interpret the data, be familiar with the data dynamics study methods and apply them to development of health strategy creation.

Healthcare facility chain and activities

Healthcare service providers

Pursuant to Act No. 578/2004 Coll. on Healthcare Providers, Healthcare Staff and Healthcare Professional Organizations, healthcare provider is a physical or legal entity providing healthcare on basis of permit or a physical individual providing healthcare on basis of license. We distinguish outpatient´s, hospital and pharmaceutical healthcare pursuant to healthcare provision form. Detail structure of healthcare provision is stated in the table No. 1.

Outpatient´s facilities	**Hospital facilities**	**Pharmaceutical facilities**
– outpatient´s dept., – one-day healthcare provi-sion facility – stationary facility, – policlinic, – home nurse care agency, – common examination and treatment facility, – mobile hospice	– hospital • general, • specialized, – therapeutic facility, – hospice, – nurse care house, - natural healing spa, – bio-medicine research facility	– in hospital pharmacies, – in public pharmacies Incl. branches, – in medical aid distribut. centers, – in public pharmacies Serving to educational purposes

Table No. 1 Healthcare facilities pursuant to Act No. 578/2004 § 34 clause 2 and Act No. 140/1998 Coll. as amended

Number of healthcare providers represents their basic statistical characteristics.

Chain

Let´s specify the meaning of healthcare facility chain. We shall interpret it as all facilities providing healthcare services in particular area and time, structured according to the above cited law. The following table provides an overview of all

Types of healthcare facilities	Number				
	PZS, ktorí prevádzkujú danýdruhzariadenia	zdravotníckychzariadení	pracovných miestsamostatnýchodbornýchzdravotníckychpracovníkov	postelí	denných miest prepacientov
Total	12180	13051	23735,10	44073	4005
Outpatient´s healthcare	9300	10060	11283,58	–	2174
General Outpatient´s department	2809	2941	2863,18	–	–
Special Outpatient´s department	5 719	6 132	6 146,55	–	–
Emergency out-Patient´s dept.	28	28	370,81	–	–
One-day health-Care provision facility	72	85	135,33	–	375
Hospital health-Care facility Incl. outpatient´s departments	173	186	9244,58	44073	1 831
general hospital	65	76	7772,12	25075	1707
specialized hospital	44	44	1220,04	6142	114
Therapeutic fac.	20	20	95,39	1446	10
hospice	9	9	29,14	156	–
Pharmaceutical care	1631	1665	3061,49	–	–
Public pharmacy	1460	1468	2968,49	–	–
Public pharmacy branch	55	55	69,59	–	–
Public pharmacy Serving to educatio-nal purposes	1	1	13,88		

Table No. 2 Number of healthcare facilities pursuant to[3] healthcare facilities. Healthcare facility chain is regulated by the Act No. 277/1994 Coll. on Healthcare as amended. The chain consists of outpatient´s departments including emergency healthcare stations and outpatient´s departments in the social service facilities, home nurse care agencies, emergency health services, specialized outpatient´s departments, dialyze stations, policlinics, independent common examination and therapeutic facilities, I., II., and III. category hospitals with policlinic, I., II., and III. category hospitals, university hospitals, university hospitals with policlinic, high-specialized medical institutions, therapeutic facilities for long-term ill, hospices (palliative care facilities), geriatric centers, psychiatric therapeutic facilities, psychiatric hospitals, psychiatric stationary facilities, drug addiction therapeutic facilities, special therapeutic facilities, natural therapeutic spas, public pharmacy branches, medical aid distribution stations and dental technical centers. [2]

National center of healthcare information published overview of healthcare chain as of Dec 31, 2011 in its Almanac 2011 [3] structured according to healthcare facility types.

The numbers published don´t say much about healthcare providers´ chain situation. To obtain better view of the situation, we should make indicators from the numbers and present them as a time line within a few years. Then we create similar system of indicators for particular regions and surrounding countries and compare them. In order to obtain the time line instead of laborious rewriting the data from NCZI almanacs, we will use data from SLOVSTAT. However, the data don´t cover year 2011, thus we will add them from NCZI Almanac (Table No. 3). Total number of beds provides the information about available sources for service provision to hospitalized patients.

We can see from the data that reduction of hospital beds count has been done in the latest years in Slovakia that mostly affected general hospitals. Number of beds in specialized hospitals has dropped less dramatically. Such dynamical development can be recognized in total counts; indicator “total beds per 100 thsd. population median” highlights the differences even more. Let´s look at the differences between the Slovak regions as provided by RegDat [4]. We chose total count of beds in healthcare facilities in 2010 and recalculated it to the population median.

Total counts document the differences between particular regions that are even more emphasized upon recalculation to 100 000 citizens. Bratislava region with the Slovak capital city disposes with the highest number of beds since there are mainly specialized healthcare facilities in Bratislava, serving to the whole country. Trnava region has the lowest count of beds per citizens´ headcount, which can be explained partly with its location nearby the Slovak capital city. On the contrary, Prešov region has the 2nd highest count of beds in absolute figures and also upon recalculation to citizens´ headcount.

	Years						
	2005	2006	2007	2008	2009	2010	2011
Population median	5387285	5391184	5397766	5406972	5418374	5431024	5398384
Healthcare facilities in total							
beds	48622	47875	47524	46742	46878	45889	44073
beds/ 100000 citizens	902,53	888,02	880,44	864,48	865,17	844,94	816,41
General hospitals							
beds	28983	28352	28328	27866	27658	27276	25075
beds/ 100000 citizens	537,99	526,28	525,83	517,25	513,39	506,3	465,45
Specialized hospitals							
beds	6908	5972	5960	6046	5990	5974	6142
beds/ 100000 citizens	128,23	110,85	110,63	112,23	111,19	110,89	114,01

Table No. 3 Beds in healthcare facilities in total and recalculated per 100 000 citizens (median).

	Population median (as of July 01)	Total beds In healthcare facilities	Total beds per 100000 citizens
Bratislava region	625834	5069	809.96
Trnava region	562391	2325	413.41
Trenčín region	599018	3089	515.68
Nitra region	705193	3942	559
Žilina region	698009	3846	551
B. Bystrica region	652800	3780	579.04
Prešov region	808532	5830	721.06
Košice region	779247	5369	689
Slovakia	5431024	33250	612.22

Table No. 4 Beds in healthcare facilities in total and recalculated per 100 000 citizens in particular Slovak regions

	Total beds in hospitals	per 100000 citizens
EU (27 countries)	2700506	538.2
Czech Republic	73746	701.0
Hungary	71818	718.2
Austria	64008	762.9
Poland	251456	658.5
Slovakia	34850	641.8

Table No. 5 Total beds in EU and neighboring countries in 2010, total count and recalculated per 100 000 citizens. Source: [5]

Looking at the table No. 5, we can see that Slovakia disposes with the lowest count of beds per capita among the neighboring countries being members of EU. Compared to the entire EU, Slovakia has more beds per 100 000 citizens as in entire EU. The table clearly demonstrates that absolute counts of beds would be misleading when comparing territories with various citizens´ headcount.

Considering OECD member countries Slovakia is also included in, the highest count of beds per capita is in Japan and Korea – more than 8 beds per 1 000 citizens in 2009. Japan and Korea have majority of hospital beds intended for long-term medicare. Number of beds in hospitals highly exceeds OECD average count also in Russia, Germany and Austria. On the other hand, big developing countries in Asia (India, Indonesia and China) have relatively few hospital beds compared to OECD countries. Total number of hospital beds per capita has slightly decreased in most OECD countries in the last decade. While in 2000 there were 5.4 hospital beds per 1 000 citizens, the count decreased in 2009 to 4.9 beds. The decrease was caused partly by medicine technology progress that allowed for change to one-day surgery, and resulted in reduced need for hospitalization. Decrease of hospital beds count was accompanied with reduced hospitalizations and average stay in hospital in many countries, except Korea, Greece and Turkey where number of beds per capital increased in years 2000 and 2009. [6]

Situation in Slovakia is characterized by the table *Bed care in specialized facilities of hospital healthcare in* NCZI Almanac. [3]

Hospitalization

Hospitalization is the basis of monitoring of activities of facilities where patients´ stays for purpose of medical treatment. We shall interpret it as every termination of hospitalization in a hospital ward in the form of release, death or relocation to other hospital ward. Hospitalization is characterized by a few indicators; most significant of them refer to headcount of hospitalized and dead patients, average time of treatment, bed utilization in days and in percentage (engagement).

Headcount of hospitalized patients

The number of beds doesn´t say much about the utilization that could be used for deduction of the need or substantiation of cost. Headcount of hospitalized patient is calculated so that we consider 1 calendar day as 1 hospitalization day in which the patient received all services provided by the facility, i.e. therapy – treatment including accommodation and catering. Both first and last day of hospitalization are considered one hospitalization day.

Average treatment time

Average Length of Stay (ALOS) refers to patient´s duration of stay in hospital / facility. It is expressed as proportion of total treatment days to hospitalized patients´ headcount. The indicator is closely correlated with bed facility capacity. In general, complicated cases require longer hospitalization and thus higher cost. This indicator used to be adjusted to Case-Mix, i.e. set of patients divided according to diagnosis category. The term is often used in relation to DRG. Average adjusted length of stay usually means that it has been adjusted according to patients´ structure.

ALOS is often used as an efficiency indicator. If the other hospitalization characteristics were the same, shorter stay would save cost and shift hospitalization care towards cheaper outpatient´s care in acute cases. On the other hand, shorter stay tends to be more care-intensity demanding and thus more costly. Too short stay in hospital can adversely affect medicare result or reduce comfort and healing of the ill. It could cause faster hospitalization recurrence and reduce the disease episode cost only slightly, or even increase it. [6]

Looking at the hospitalization statistics, we can see that children are those most frequently hospitalized. Most death cases occur at the internal medicine wards, reflecting the absence of long-term professional care in Slovakia, therefore older patients are usually hospitalized at these wards. This significantly contributes to healthcare cost increase and corresponding higher count of days of hospitalization. On the other hand, patients spend on bed at the internal medicine ward only 7 days in average while they spend approx. 27 days when hospitalized at psychiatric ward.

Long hospitalization has been reported also at pneumology and phtiseology wards where mostly older patients stay. (Table No. 6).

Special department	Hospitalized patients		Dead	
	count	per 10 000 citizens	count	per 1 000 hospitalized
In total	1006820	1863	28084	27,9
internal medicine	137645	254,7	7206	52,4
infectology	15098	27,9	75	5
pneumology phtiseology	23524	43,5	770	32,7
neurology	59951	110,9	1022	17
psychiatry	32930	60,9	146	4,4
pediatrics	93669	919,4	52	0,6
gynecology and neonatory	124549	449,2	49	0,4
surgery	117870	218,1	1420	12

Special department	Total days Of treatment	Average days Of treatment	Bed utilization In days	Bed utilization in %
In total	8074236	8	237,7	67,7
internal medicine	958831	7	261,7	72,6
infectology	92336	6,1	168,7	54,5
pneumology phtiseology	307144	13,1	188,1	58,5
neurology	427705	7,1	258,6	71,9
psychiatry	893505	27,1	287,3	79,7
pediatrics	421485	4,5	237,8	66,2
gynecology and neonatory	579574	4,7	218,7	60,9
surgery	595447	5,1	226,4	63,7

Table No. 6 Hospitalizations at selected special departments in Slovakia in 2011. Source: [3]

$$\text{Average treatment time} = \frac{\textit{total days of treatment}}{\textit{headcount of the hospitalized}}$$

Bed utilization in days states how many days was a bed engaged during the monitored period:

$$\text{Bed utilization in days} = \frac{\textit{average count of beds}}{\textit{total treatment days}}$$

Bed utilization in % for certain period is calculated analogically. The following is formula for calculation during a calendar day.

$$\text{Bed utilization in \%} = \frac{\textit{average count of beds* 100}}{\textit{total treatment days}}$$

You can see on the table that the highest beds utilization has been reported at psychiatric wards and the lowest utilization at infectious wards. Hospitals often collect more data than stated herein. Those interested in detail information on hospitalization ward activities can consult it for example with ÚZIŠ ČR publication [7], or with IFHIMA (International Association of Healthcare Informatics and Management) textbooks [8]. This type of information is useful mainly for planning of healthcare funds allocation. Modeling of particular wards activity within healthcare facility chain can reveal hidden reserves and the need for adjustments. The publication authors cooperated on the application of the British model of maximum adjustment to the conditions in former Czechoslovakia at the end of the 70s. [9] Considering the then situation, the attempts didn´t find response in wider professional public.

Statistics of outpatient´s care covers services provided by healthcare staff without hospitalization. Outpatient´s care facilities refer to: (pursuant to Act No. 578/2004) outpatient´s department, one-day medicare provision facility, stationary ward, policlinic, home nurse care agency, joint facility of examination and therapeutic units, mobile hospice. Look closer at the statistics of most frequently operated outpatient´s departments, namely general practitioner, general practitioner for children and adolescents, Emergency Room (ER) and dental outpatient´s departments. In these facilities, number of outpatient´s departments is monitored according to specialization, jobs count (type of job) and frequency of visits/ treatments.

Treatment

Treatment (examination) represents a system of all acts performed by physician or nurse at a single patient in the same outpatient´s department (workplace, home of the sick person) during a single patient´s visit.

Department specialization	Number of		
	departments	jobs – in total	Visits in facility Or out of facility
In total	15251	27070,92	65904572
general practitioner for children and adolescents	1141	2065,17	7011578
general medicine	2129	4091,08	17332713
ER – general outpatient´s care of the adult In the facility	77	326,70	361969
ER – general outpatient´s care of the adult Out of facility	64	38,13	81265
ER – general outpatient´s care of the children And adolescents - in the facility	58	211,81	233550
ER – general outpatient´s care of the children And adolescents – out of the facility	24	12,70	11583
ER – general dental care of the adult	43	40,38	76201

Table No. 7 Total departments, jobs and visits in general outpatient´s care facilities in Slovakia in 2011. Source: [3]

Table No. 7 shows number of departments, jobs and visits in general outpatient´s care facilities in Slovakia.

Most facilities are general practitioners, followed by general practitioners for children and adolescents – less by approx. 50%. Statistical almanac offers overview per particular regions. OECD uses data recalculation per headcount of citizens in relevant population [6], which provides for better interpretation and comparison. For example, number of consultations per capita in Slovakia in 2011 reached 12.2, similar to Hungary (12), and the Czech Republic (11.2), while number of consultations was much smaller in Austria (6.9) and Poland (6.8).

The extent of healthcare services is much higher than described in this chapter. The statistics develops according to healthcare service system development. Detail figures within the system and the activities, associated with satisfaction of consumers´ expectations have been ever more a subject of planning and creation of strategies in the area of healthcare services.

Healthcare staff and education system

People represent a natural part of the healthcare service resources; we often speak of human resources. They refer to one of major factors influencing the quality of provided healthcare. Healthcare systems performance depends mainly on knowledge, skills and motivation of persons responsible for the service provision.

Human resources usually represent the biggest item in the healthcare service cost. Labor cost in many countries represents two thirds of total common expenditures, or even more. [10] Majority of healthcare system staff is expected to pass demanding and long-term or never-ending special preparation. Thus, healthcare education system is an important part of human resources development. In this chapter, we will discuss the characteristics that are routinely processed within healthcare statistics. Employee headcount refers to the basis but the headcount only would be misleading since people can work shorter than monitored period, can be enrolled in part-time engagement and are sometimes on sick leave. Thus, an indicator was introduced, recalculating the employee headcount to actually worked out load.

Recalculated employee headcount

Recalculated employee headcount (physicians and nurses) is a sum of employee workload (physicians and nurses) with concluded regular labor contract at healthcare facility, recalculated according to weekly working time at particular facility. Usual weekly working time refers to 40 hours per a week or 33.5 hours a week at hazardous workplaces. Healthcare staff is divided in healthcare profession categories regulated by Act No. 95/2004 Coll. *On the Terms of Acquisition and Acceptation of Professional Qualification and Special Professional Qualification to Perform as a Physician, Dentist and Pharmacist as amended.*

We stated employee headcount at outpatient´s healthcare facilities. (Table No. 7) Employee headcount is recorded according to the regional facility founder according to gender, occupation and age. Statistics of physicians and dentists is recorded separately, as well as for nurses, birth assistants, according to founder, education degree, age category and gender. Headcount of students of high schools of medicine and faculties of medicine is recorded. The following table shows employee headcount according to selected categories (Table No. 8).

Headcount of physicians was interpreted in the past in total together with the count of hospitals and other healthcare facilities as a proof of the country population healthcare quality. Thus, the count of physicians and hospitals was significantly increased in the 60s and 70s of the previous century. Later it was realized that population health condition improvement wasn't priority, comparable to the Western Europe or USA population. Today, more attention has been paid to health condition improvement than to some figures. According to OECD, number of physicians per 1000 citizens had been increased within 2000 – 2009 by 1.7% in average per a year in most of OECD countries. The growth trend was very fast in the countries with fewer physicians in year 2000 (Turkey, Chile, South Korea, and Mexico) as well as in UK

Registered employee headcount in persons							
	In total	**Health staff in total**	**Physicians**	**Nurses**	**Birth assistants**	**Other staff**	**Non-health in total**
Slovak Republic	105 743	78 842	17 849	32 043	1 837	1 400	25 457
Recalculated employee headcount per 100 000 citizens							
Slovak Republic	1 956,64	1 458,87	330,27	592,91	33,99	25,91	471,05

Table No. 8 Employee headcounts at healthcare facilities according t selected categories in Slovakia in 2011. Source: [3]

and Greece. In United Kingdom of Great Britain, headcount of medicine program graduates was above the OECD average in this period of time, which resulted in high and rising count of physicians. On the other hand, no increase in the count of physicians per capita was reported in Estonia, France, Israel and Poland while significant decrease was reported in Slovakia. Such decrease could be attributed to partial decrease of headcount of health specializations graduates since the end of 1990. Headcount of physicians started decreasing in France in 2006 in relation to reduced count of new students enrolled at the high schools/ faculties of medicine during years 1980 and 1990. [6]

Economic indicators

Finally, look at the healthcare facilities economic performance since money also represents healthcare sources. This section is dedicated to the micro-economic data characterizing healthcare sources. Separate chapter here discusses the macro-economy and its relation to the public health. At first, we have to define at least two terms related to corporate economy, since majority of healthcare providers are business subjects nowadays. Corporate economy is characterized with cost, revenues and performance (profit/ loss). Taking in account the nature of this publication, reader can find many textbooks dealing with this topic, for example Lisý et al.[11] in Slovak language or classicist Samuelson et al. [12] in English language.

Cost

Cost in pecuniary expression represents consumption of material and work. It is an amount expressed in money that must be spent to reach revenues. Cost represents one of fundamental indicators of economic effectiveness that can be simply described with words “how to provide the best/ highest performance at the lowest possible price”. We know cost associated with procurement of materials required for equipment operation, for example instruments, cars, ambulance cars, buildings, employees, material, pharmaceuticals, etc. Operation requires operating cost, i.e. services associated with building maintenance or employee training. Statistical Almanac NCZI states *Cost of Treatment Days in Euro and Points Allocated in Selected Healthcare Facilities in Slovakia and Total Cost of Provided Healthcare.*

Revenues

Every healthcare institute is a company from economic point of view, paid from various sources to provide services, sale of pharmaceuticals, technique, education or transport; or often for combination of the stated services. Taking in account business nature of contemporary healthcare system, revenues include also non-operating revenues, for example rental payment for premises or income from deposits. Sales from sold goods or provided services represent major part of the revenues. The revenues should at least cover the cost in common business sense.

Performance result (profit/ loss)

Performance result (profit, loss) represents the difference between total cost and total revenues. Discussion has been held in the healthcare sector whether healthcare services providing facilities should or shouldn´t create profit. The fact that the state-owned (in part or fully) facilities create mostly negative profit or loss, has been discussed in significantly lesser extent. Let´s leave the attitude to profit on the politicians and focus on the table provided by NCZI and containing data about cost, revenues and performance in Euro in healthcare facilities according to founder and legal status, healthcare facility type and site.

Reader should study the data applicable to Slovakia in NCZI publications. OECD published healthcare economy data in its DB and periodical almanacs.

National Health Accounts

Objectives

At the end, we will briefly pay attention to the specific method of healthcare services financing expression through the national heath accounts. National accounts provide the data within a broad spectrum of inland economic activities as well as transactions with foreign countries. The methodology has been elaborated since the beginning of the 40s of the previous century when the attempt was made in Great Britain for estimation of war economy sources.

UN proposed model in 1952 (System of national Accounts – SNA) was the first international system. Creation of the European System of Accounts (ESA) in 1970 represented a significant step ahead. [13] Health accounts represent a separate chapter within the system with the last revision effective since 2011. [14]

Health accounts provide systematic description of cash flow related to healthcare material and services consumption. Health accounts are ever more expected to provide the ideas of improvement (together with statistical information) of healthcare system evaluation and monitoring analytic tools. One of the priorities is to create reliable and timely data comparable with other countries in real time. It is necessary for monitoring of healthcare expenditures and driving forces development, and further for comparison with other countries. Health accounts are used in two major directions: on international level with focus laid on selection of internationally comparable expenditures data, and on national level focused on detail healthcare expenditure analysis, and mainly on time comparison.

Better said, Health Account System 2011 is aimed at specifying the frame of data suitable for international comparison of healthcare expenditures and healthcare system analysis, and at providing the tool supporting the healthcare system monitoring and analysis that could be extended to other countries. Last but not least, the system provides definition of internationally harmonized boundaries of healthcare serving to monitoring of expenditures.

Classification

Classification of the system items represents a major contribution to standardization of healthcare services recognition and comparison from economic point of view. There are three basic classifications: healthcare providers, functions and health services financing. Additional indicators refer to: classification of types of revenues from healthcare funding systems and classification of healthcare provision factors. Moreover it includes classification of recipients according to age, gender, disease, social-economic characteristics or region; as well as classification of human resources, healthcare goods and services.

Examples of use

Information from national health accounts was used by health strategy authors and scientists in various ways. Newhaus used the information to study the factors that mostly influence healthcare cost in USA and proved that technology change refers to most important factor. [15] Another frequent way of data utilization from national health accounts is assessment of ageing impact on healthcare expenditures. [16] Data from national health accounts were used also for the analysis of public healthcare services funding in Europe, [17] Africa [18], China [19] and other countries. Unfortunately, we didn´t find any publication dealing with this topic in Slovakia.

Summary

In this chapter, we discussed basic statistics of healthcare sources, described healthcare providers chain and characteristics of their activities. We stated basic opinions about human resources in healthcare services and discussed finances in the terms of cost and revenues. Finally, we briefly described the fundamentals of national health accounts and their use. We believe that reader will get overview of the subject matter and be inspired by the examples of research in the world that could be applied also in Slovakia.

References

1. WHO. *Guide to producing national health accounts: with special applications for low-income and middle-income countries*. 2003, World Health Organization.
2. SOURCE: STATISTICAL OFFICE*: Selected indicators, methodical explanations– healthcare sector*. 2013 [cited 2013 4.9.]; Available from: http://portal.statistics.sk/showdoc.do?docid=4325.
3. NCZI. *Healthcare Almanac of the Slovak Republic 2011*. 2012, National Center of Healthcare Information: Bratislava. p. 256.
4. Source: Statistical Office of the Slovak Republic. *Database of Regional Statistics RegDat*. 2013 [cited 2013 22.5.]; Available from: http://px-web.statistics.sk/PXWebSlovak/.
5. EUROSTAT. *EUROSTAT-DATABASE*. 2013 [cited 2013 19.02.]; Available from: http://epp. Eurostat.ec.europa.eu/portal/page/portal/statistics/search_database.
6. OECD. *Health at a Glance 2011*. 2012: OECD Publishing. 199.
7. UZIŠ ČR. Hospitalization Care 2012. Healthcare Statistics. 2013 Institute of Healthcare Information and Statistics of the Czech Republic: Praha. p. 142.
8. IFHIMA. *Education Module for Health Record Practice. Module 4 – Healthcare Statistics*. 2013 [cited 2013 7.9.]; Available from: http://www.ifhima.org/ed_modules/Updated 2012/Module 4 Healthcare Statistics.pdf.
9. Aspden, P., Mayhew, L., Rusnak, M. *DRAM: A Model of Health Care Resource Allocation in Czechoslovakia.* OMEGA, 1981. **9**(5): p. 509 – 518.
10. WHO. *The world health report 2000 — Health systems: improving performance.* 2000, Geneva: World Health Organization.
11. Lisý, J. a kol. *Economy in New Economy* 2007: IURA EDITION.
12. Samuelson, P., Nordhaus, W. Economics. 19th ed. 2009: McGraw-Hill/Irwin.
13. EUROSTAT. EUROPEAN SYSTEM OF ACCOUNTS, ESA 1995. 1995: Brussels.

14. OECD, EU, WHO. A System of Health Accounts. 2011 ed. 2011: OECD Publishing.
15. Newhouse, J. P. Medical care costs: how much welfare loss? The journal of economic perspectives: a journal of the American Economic Association, 1992. 6(3): p. 3 – 21.
16. Gerdtham, U. G. et al. An econometric analysis of health care expenditure: A cross-section study of the OECD countries. Journal of Health Economics, 1992. 11(1): p. 63 – 84.
17. Rechel, B., Brand, H., McKee, M. Financing public health in Europe. Gesundheitswesen, 2013. 75(5): p. e28 – e33.
18. Sambo, L. G., Kirigia, J. M., and Orem, J. N. Health financing in the African Region: 2000 – 2009 data analysis. International Archives of Medicine, 2013. 6(1).
19. Long, Q. et al. Changes in health expenditures in China in 2000s: Has the health system reform improved affordability. International Journal for Equity in Health, 2013. 12(1).

Chapter 11

Health and healthcare economic indicators

Chapter objectives

Slovak citizens have witnessed many changes during recent years that were reflected in their life and health. Differences in income, property and health condition have been more and more apparent. Factors as nutrition and hygiene improvement, innovations in medical technologies and healthcare infrastructure gradually prolong human life. Relative effect of these factors depends on the economic development, marked by some authors as the most important one. We could discuss the extent of economic development effect but it´s true that we often cannot avoid speaking of economic backgrounds of an individual or population, which allows for utilization of healthcare services that directly or indirectly influence the life of an individual or the whole population. Concurrently, we have to think about the opposite relation – how health influences the economic development. [1] Healthier workers are physically and psychically more fit and robust. Their production rate is higher, they earn more and it is less likely that they will be sick/ draw sick leave.

It is apparent from such a brief introduction that mutual relations should be studied. It is also necessary to consider another measure, namely the price of health. We could simply state that we buy health in various ways: healthy lifestyle used to be more costly than that of a person who doesn´t care. We pay for prevention exams, therapy and life maintaining. Price has important meaning in the considerations of health.

This chapter is aimed at presenting most frequently used economic indicators and documenting their application in the population health studies on the examples. We will discuss the macro-economic indicators on the state and regional level. Healthcare economic parameters are discussed in the chapter about healthcare services.

Basic macroeconomic indicators

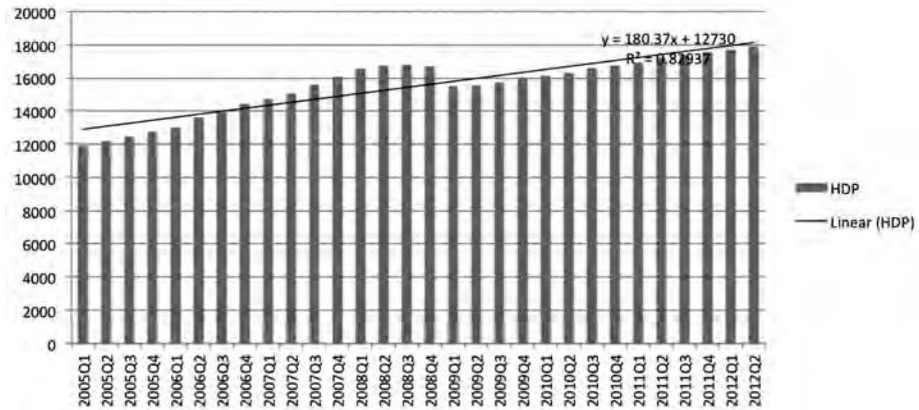
Methodology of such and other indicators preparation is described in detail in the EU System of National and Regional Accounts in EC [2] and is fully consistent with revised globally used methodic – System of National Accounts [3]. Prior to presentation of such systems, we will discuss basic macro-economic indicators. We can consider the two basic documents and other, derived from them, as a method of data standardization.

Development rate of a country national economy is evaluated on basis of macro-economic indicators. They include gross and net domestic product (GDP and NDP), gross and net national product and national income. Moreover, other indicators have been used for measurement of economy, so called alternative indicators of gross national product. It includes net economic wealth and index of human development.

Gross domestic product (GDP)

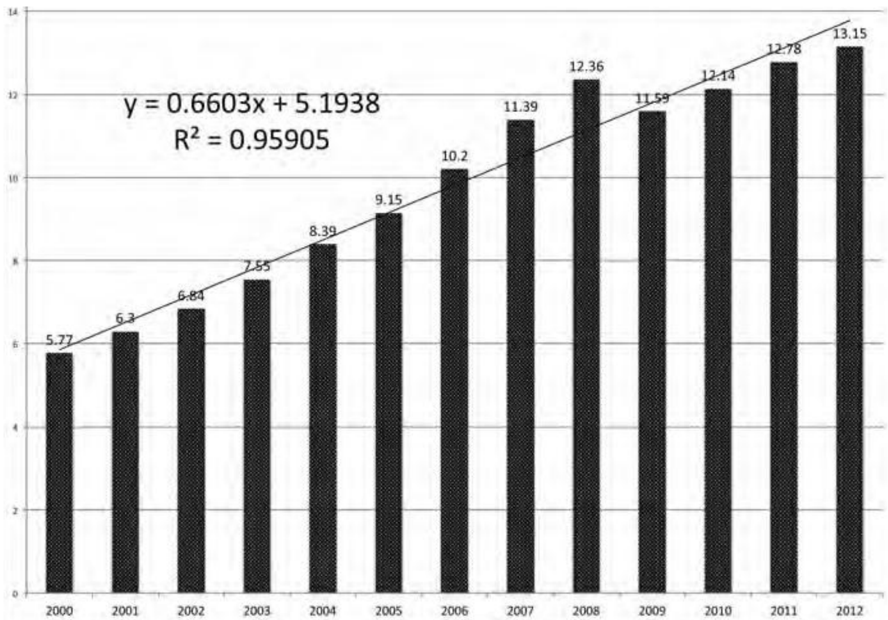
GDP is a macro-economic indicator of particular country economic performance. GDP is a summary of produced goods and provided services created by production factors in the country territory per one year, regardless the production factor owners and their citizenship. Thus, GDP includes the value of final goods and services. From the value point of view, GDP represents market value of all final property produced in the country territory. [4, 5] GDP indicator is used by majority of world countries and is based on commonly available statistical data.

To explain the essence of the combined indicator, we should point out the parameters used for GDP calculation. Expenditure method is the one serving for such calculation. It includes expenses of a household for common consumption (food, clothes, boots, drugstore goods, furniture, refrigerator, washing machine, cars, rental payment, gas, power, water and medical payments),



Picture No. 1 Slovak GDP per quarters during years 2005 – 2012. Source: SLOVSTAT

private gross domestic investments of companies (funds spent for building of family houses and apartments, for purchase of capital assets, for inventories), state expenditures for purchase of the goods and services (those bought from private sector, expenditures for building of roads and highways, expenditures for security, education and healthcare, science and research, legislation, etc.) and net export. Moreover, other GDP calculation methods are used, for example based on income or added value of



Picture No. 2 Slovak GDP recalculated per capita in b.c. (thsd. EUR). Source: SLOVSTAT particular producers and indirect taxes.

GDP development in Slovakia in particular quarters for 7 years is illustrated on the Picture No. 1. GDP decrease was reported at the end of 2008, followed with gradual increase. In the last quarter, GDP exceeded its value before the decrease. Regression line confirms the growth and its trend is very good ($r^2 = 0.8$). GDP recalculation per capita shows trend similar to that applicable to the whole country (Picture No. 2).

But trend itself is not enough for derivation of other conclusions. We should look at Slovakia position among the neighboring countries and decide in which category of economic performance Slovakia should be included.

Country	GDP per capita	
	Eurostat 2012	World Bank 2011
EU 27 members	23 200 €	
Czech Republic	11 400 €	18 700 $
Hungary	8 800 €	12 730 $
Austria	32 100 €	48 170 $
Poland	8 500 €	12 380 $
Ukraine		3 130 $
Slovakia	9 400 €	16 190 $

Table No. 1 GDP per capita in 2012. Comparison of EU average and neighboring countries.
Source: Eurostat 2012 and World Bank 2011

Comparison with the neighboring countries and EU average shows that Slovakia doesn't rank among wealthy countries of EU. Austria has approx. 3 times the Slovak GDP per capita and EU average is approx. the double. According to the World Bank table, GDP limit in countries with low income refers to USD 571 and less per capita while in the countries with high income it is USD 38,661 and more. Thus, Slovakia ranks among the countries with medium income level, similar to majority of its neighbors except Austria.

We will state a few examples of GDP use for the study of the relation between economy and factors affecting or characterizing health. The study of the relation between EU countries economic strength and health condition showed statistically remarkable dependence of GDP per capita and total healthcare cost per capita. As well, relation between GDP per capita and population health condition indicators was confirmed as statistically significant. [6] There is a logical question in what extent could this fact be used in particular country, since we know the differences between particular countries not only in the area of GDP but also in the life expectancy. Further examples can be found in foreign publications. [7 – 9]

Example

As an example, we will try to find an answer to the question whether the differences in GDP between Slovak regions are statistically significant in relation to the average years lived if born in particular region. In the regional statistical DB (hereinafter “RegDat”), [10] containing the time lines of economic and social-economic development indicators in the Slovak regions (districts, regions and zones – NUts2 according to data availability), we searched for the following indicators:

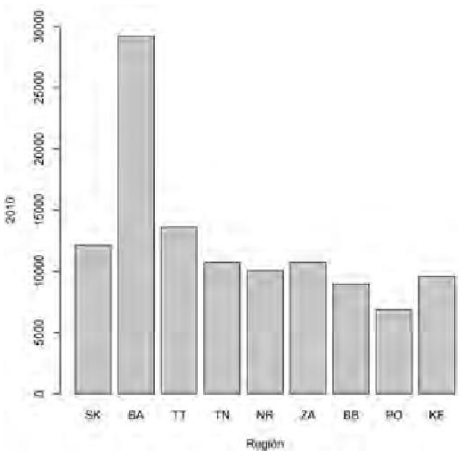
- regional GDP (regular prices in €) based on territory, indicator type and year, for period 1995 -2010;
- citizens´ headcount – population median (as of July 01) based on 5-year category of age, region, gender and year, for period 2001 - 2011;
- Life median at birth based on territory, gender and year, for years 2001 - 2011.

Since the population median and average life expectancy at birth are included in DB only for period 2001- 2011, but GDP data are recorded only till 2010, we will work with years 2001 through 2010. All data were entered in the Excel format. To settle the population difference in particular regions, we calculated *regional GDP per capita* as a proportion of two indicators – regional GDP (with applied criterion of workplace based preparation) and average citizens´ headcount in particular region (based on the residence principle). Comparison of the two indicators based on different principles doesn´t cause any major problems in majority of regions. The indicator is overestimated in the regions with high rate of traveling to work from the surrounding regions (mainly capital city region). [11] We will ignore this warning in our example. Thus, the first step will refer to transposition of relevant data in {R} environment. Since we entered data from RegDat website in Excel, their reading in the variables shouldn´t be complicated (Picture No. 3). We will create the following variables: *LE_M* for average life duration of men, *LE_Z* for average life duration of women, *POP_M* a *POP_Z* for population median, *GDP* , and *GDP_obyv* for recalculated GDP per capita. The result can be plotted in the column chat separately per each year or region. As an example, we chose year 2010 and Bratislava region. Process of plotting in the column chart is on the picture No. 3.

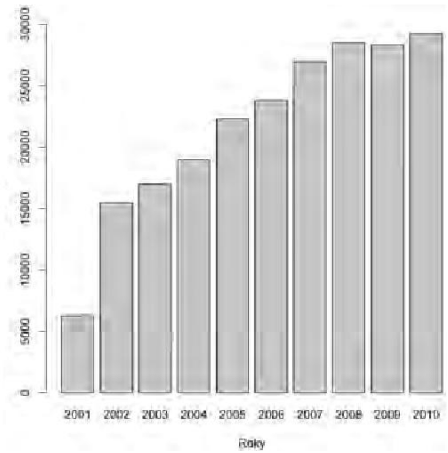
```
> LE_M <- read.table(„clipboard“, header=TRUE, sep=““, na.strings=“NA“, dec=“,“, strip.white=TRUE)
#reading variables from Excel table to the frame
> LE_M
region X2001 X2002 X2003 X2004 X2005 X2006 X2007 X2008 X2009 X2010
1   SR 69.51 69.77 69.77 70.29 70.11 70.40 70.51 70.85 71.27 71.62
2   BA 71.48 71.70 72.03 72.05 72.04 72.12 72.47 72.96 73.37 73.55
```

```
3   TT 69.49 69.69 69.81 69.99 70.37 70.59 70.77 70.96 71.38 71.74
4   TN 70.59 70.91 70.89 71.11 71.08 71.50 71.42 71.78 71.85 72.36
5   NR 68.79 69.17 69.28 69.63 69.83 70.00 70.06 70.06 70.41 70.76
6   ZA 68.89 69.26 69.65 70.00 70.00 70.09 70.06 70.12 70.38 70.76
7   BB 67.99 68.25 68.46 68.80 68.90 69.25 69.41 69.76 70.03 70.41
8   PO 69.68 69.86 70.08 70.38 70.48 70.61 70.65 70.74 71.14 71.43
9   KE 68.34 68.51 68.73 68.87 69.07 69.22 69.35 69.67 70.11 70.54
> attach(LE_M) # creation of separate variables from the frame
```

Picture No. 3 Process of data transposition in {R} environment, - average life duration



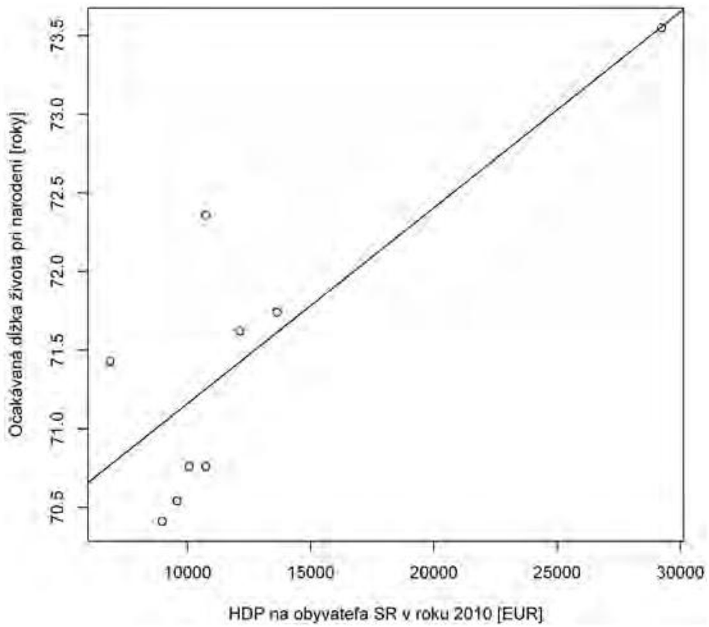
> *barplot(HDP_obyv$X2010, names.arg=c("SK","BA","TT","TN","NR","ZA", "BB", "PO", "KE"), xlab="Región", ylab="2010", ylim=c(0, 30000))*



> *SR=c(HDP_obyv, HDP_obyv[2,3], HDP_obyv[2,4], HDP_obyv[2,5], HDP_obyv[2,6], HDP_obyv[2,7], HDP_obyv[2,8], HDP_obyv[2,9], HDP_obyv[2,10], HDP_obyv[2,11]) SR > barplot(SR, names.arg=c("2001","2002","2003", "2004", "2005", "2006", "2007", "2008", "2009", "2010"), xlab="Roky", ylab="Bratislava", ylim=c(0, 30000))*

Picture No. 4 Illustration of GDP per capita in 2010 according to regions, GDP per capita in Bratislava region in particular years.

If we arrange the data according to GDP per capita and make linear regression between GDP per capita and life expectancy at birth at men living in 2010 in particular regions, we will get not so strong linear dependence. Regression coefficient 0.6 isn´t very convincing despite of statistically significant regression coefficient (Picture No. 5). Nevertheless, we cannot refuse the hypothesis of the existing relation between the two parameters; anyway we should take in account other circumstances at its interpretation



> le <- LE_M$X2010
> hdp <- HDP_obyv$X2010
> reg <- c(„SR“,“BA“,“TT“,“TN“,“NR“,“ZA“,“BB“,“PO“,“KE“)
> mm_10 <- cbind(hdp, le)
> srt <- mm_10[order(hdp),]
> vzth <- lm(srt[,1] ~ srt[,2])
> summary(vzth)
> plot(srt[,1],srt[,2], xlab=(„GDP per capita in Slovakia in 2010 [EUR]“), ylab=(„Life expectancy at birth[years]“))
> abline(lm(srt ~ srt[,1]), col=“black“)
Call:
lm(formula = srt[, 1] ~ srt[, 2])
Residuals:
Min 1Q Median 3Q Max
-6409 -1136 1330 1998 5832
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -363000 103117 -3.520 0.00972 **
srt[, 2] 5254 1443 3.641 0.00827 **
Signif. codes: 0 ‚***‘ 0.001 ‚**‘ 0.01 ‚*‘ 0.05 ‚.‘ 0.1 ‚ ‚ 1
Residual standard error: 4135 on 7 degrees of freedom
Multiple R-squared: 0.6545, Adjusted R-squared: 0.6051
F-statistic: 13.26 on 1 and 7 DF, p-value: 0.008273

Picture No. 5 Illustration of relation between GDP per capita and life expectancy in Slovakia in 2010. Data source: SLOVSTAT that we didn´t consider in this case. The fact is that GDP per capita isn´t evenly distributed within a region; mainly large companies´ managers earn much. In many cases, workers who generate GDP don´t have residence in particular region but travel for work, etc. It is necessary to consider what the life expectancy means and interpretation limitations of this parameter. This was already discussed in the respective chapter.

GDP per capita

GDP represents an aggregate value and doesn´t specify the income distribution in a country, nor does it point out to the growth resulting from increased defense or police cost, or cost of health and education. Therefore, we have been continuously seeking an indicator that would provide us with more detail overview of economic situation and the conditions it has evolved in. GDP per capita indicator is used for comparison of economic performance of countries. GDP calculation per capita is made as proportion of nominal GDP to the population of particular country.

PPP

Despite of GDP per capita calculation, there are many issues that remained unsolved by the economists. One of them refers to PPP - Purchasing Power Parity. Eurostat defines it as the quantity of the goods and services that can be purchased for certain amount of money. For example, if a consumer pays today Eur 1 for a bread loaf and tomorrow the price rises to Eur 1.10, then he can buy less bread tomorrow. Thus, the value of Eur 1 dropped simultaneously with the consumer´s PPP as a result of bread price inflation. [12] Another example: if a price of hamburger in France refers to Eur 2.84 and in USA it is USD 2.20, then hamburger PPP between France and USA refers to Eur 2.84/ USD 2.20 or Eur 1.29/ USD 1. In other words, Eur 1.29 must be spent in France per each USD 1 spent in USA for a hamburger in order to get the same quantity and quality. What applies to hamburger doesn´t necessarily have to apply to other commodities; thus hamburger-related PPP isn´t equal to the official exchange rate of a currency. PPP is used for conversion of national expenditures according to product categories and GDP in various countries. Final expenditures are also called the “real ones” since, as we explained above, they are evaluated on the uniform price level within the process of conversion to a common currency, expressing thereby only the differences in quantity purchased in particular countries. Considering these features, PPP is preferred to GDP in international comparisons.

Gross National Income (GNI)

Gross national Income, GNI, is a summary of all final products and services made anywhere in the world, using production factors (i.e. land, work or capital), owned by organizations residing in the monitored territory (i.e. using so called national production factors in the monitored territory) during the monitored period. In practice, GDP and GNI are often considered equal and the difference is neglected.

Health cost indicators

Total healthcare cost (or expenditures) refers to the sum of public and private cost of healthcare. This applies to healthcare services provision (preventive and therapeutic services), parenthood planning services, nutrition and health-related aid in case of extraordinary events; but it doesn´t include water supplies and sewage. There isn´t any exact measurement of basic health expenditures in all countries. Even in countries with advanced statistical systems established, it is necessary to make some cost adjustments in order to improve the sources and methods. As we already stated, the indicators are collected by worldwide used System of National Accounts. [3] Of course, we are interested in the health related area.

WHO keeps the Global Health Expenditure Database - GHED. In this way, they have been mediating internationally comparable figures on health national expenditures for the last decade. The data have been updated on annual basis according to public-available reports (national health account reports, reports of Ministries of Treasury, Central Bank, National Statistical Office, World Bank and International Currency Fund reports and public expenditure information, etc.). Collected data are arranged in tables according to countries and indicators, and published in the form of Annual Reports [13], also available as an interactive DB. [14] Detail description of the indicators is contained on website *WHO Indicator and Measurement Registry (IMR)*.[15]

We will illustrate selected economic health-related indicators on the example of the latest WHO Global Statistical Data Report issued in 2013[13], which are derived from those stated above. The first applied indicator - *Total expenditure on health as a percentage of gross domestic product* – is a basic indicator of health financing systems. It provides information on the level of health sources compared to the country wealth, and is obtained from the System of National Accounts through synthesis of healthcare system cash flow.

Total Health Expenditures represent the sum of all financial agents that control funds for purchase of healthcare goods and services. They are subsequently recalculated per GDP part. Another indicator used - *Per capita total expenditure on health (PPP int. $)*

Notice that they are expressed in PPP, and thus comparable between countries. Finally, we state governmental expenditures for coverage of population health needs, called *per capita government expenditure on health (PPP int. $)*. This indicator contributes to the understanding of proportional level of public healthcare expenditures, recalculated per population and expressed in international USD, enabling international comparison. The following table contains all 3 indicators in Slovakia and neighboring countries (Table No. 2).

	Total health Expenditure as % GDP		**Total health expenditure per capita (PPP int. $)**		**Governmental health Expenditure per capita (PPP int. $)**	
	2000	2010	2000	2010	2000	2010
Slovak Republic	5.5	9.0	604	2097	540	1352
Czech Republic	6.3	7.5	982	1885	887	1579
Hungary	7.2	7.8	853	1601	603	1037
Poland	5.5	7.0	584	1377	409	987
Austria	10.0	11.0	2898	4398	2192	3351
Ukraine	5.6	7.8	184	527	96	298

Table No. 2 Comparison of health expenditure between Slovakia and neighboring countries during period 2000 -2010. Source: [13]

Index of human development

Recognizing the society and country development degree isn´t naturally limited by economic measures. We know that economy and human development degree is closely correlated but requires the system of complex indicators capturing broader correlations of human development, in order to identify wider aspects. *Human Development Index (HDI)* is a compound indicator that has been used since 1990 for complex comparison of economic performance of particular countries. It reflects the life expectancy, literacy and impact on maintaining worthy lifestyle. It captures various dimensions of human life conditions. [16] In its report on human development in Slovakia [17], the organization called Economic Development Center states the calculation of the index for Slovakia based on the data from period of years 2000 and 2001. HDI is calculated as combination of life expectancy, literacy, reached education and GDP per capita rates.

The first measure refers to life expectancy at birth (as an index reflecting population health and life duration). We discussed the measure in sufficient extent in a separate chapter. Percentage of citizens older than 14 years of age who are able to fully read and write a brief text about their daily life refers to the adult population literacy rate. The rate isn´t calculated on regular basis and the survey conducted on the representative group has not been made in Slovakia yet. Thus, we used to draw the data from similar survey conducted in the Czech Republic. OECD intends to conduct such survey in all EU member countries under the name *Program for the International Assessment of Adult Competencies – PIAAC* on a group of 5 thousand individuals from each country at the end of 2013. [18] The measure has 2/3 weight in the calculation. Another education measure refers to combined measure of elementary, high and university education measured as a headcount of students enrolled in certain education level regardless the age, as a percentage of total population headcount in official education age for particular education level. Combined gross rate of enrollment in the elementary, high and university education facilities represents headcount of students at all education levels as a percentage of population in the official school age for particular levels (in Slovakia, the rate is recalculated per citizens of age 6 – 22). This measure is assigned with 1/3 weight. And finally real GDP per capita (PPP$) is added in the calculation. It is calculated as a GDP per capita in domestic currency and purchasing power parity, PPP. HDI varies within 0 – 1.

To illustrate it, we state calculated HDI values for Slovakia pursuant to UNDP (Table No. 3). [19] Careful reader can notice significant increase of all indicators serving for calculation of the index, which has occurred during the last 30 years.

	Life expectancy At birth	**Education years expectancy**	**Average education Duration in y.**	**GNI Per capita (2005 PPP$)**	**Index Of human Devel. HDI**
1980	70.6		10.1		
1985	70.8		10.4	12,121	
1990	71.2	11.8	10.6	12,695	0.754
1995	72.1	12.0	11.2	10,869	0.759
2000	73.3	13.1	11.2	12,660	0.785
2005	74.3	13.9	11.6	15,720	0.814
2010	75.2	14.7	11.6	18,924	0.836
2011	75.4	14.7	11.6	19,209	0.838
2012	75.6	14.7	11.6	19,696	0.840

Table No. 3 HDI development in Slovakia during last 30 years. Source: UNDP [19]

HDI increase by approx. 60% per capita is significant during the monitored period and documents economic development of Slovakia. Increased life expectancy at birth by 5 years is favorable as well. Certain progress in the area of education has been reported. The Slovak HDI for 2012 reached 0.840, ranking our country to the 35th place amongst 187 evaluated countries. HDI increased from 0.754 to 0.840 during period of years 1990 – 2012, referring to 11% increase and average annual increase by 0.5%.

	HDI	**HDI rank**	**Life expectancy at birth**	**Education years expectancy**	**Average education Duration in y.**	**GNI per capita (PPP US$)**
Slovak Republic	0,840	35	75,6	14,7	11,6	19,696
Czech Republic	0,873	28	77,8	15,3	12,3	22,067
Hungary	0,831	37	74,6	15,3	11,7	16,088
Austria	0,895	18	81	10,8	15,3	36,438
Poland	0,821	39	76,3	10,0	15,2	17,776
Ukraine	0,740	78	68,8	11,3	14,8	6,428
Europe and Central Asia	0,771	—	71,5	13,7	10,4	12,243
Very High HDI	0,905	—	80,1	16,3	11,5	33,391

Table No. 4 Comparison with neighboring countries in 2012. Source: UNDP[19]

Comparison with neighboring countries provides even clearer picture of the situation characterizing Human Development Index. Slovakia ranked next to Hungary but the Czech Republic topped it in all indicators. However, the V4 countries significantly differ from Austria with more than double GNI per capita compared to Hungary, and much higher values than all other V4 countries. On the other hand, Ukraine lags behind significantly. Except Ukraine, all countries in the table No. 4 ranked above the Europe and Central Asia average. Austria is also behind 14 countries with the highest HDI value (Norway, Australia, USA, the Netherlands, Germany, New Zealand, Ireland, Sweden, Switzerland, Japan, Canada, South Korea, Hong Kong and Island).

Summary

In this chapter, we outlined the basic health macro-economic opinions. Considering the need for deeper economic knowledge, we could discuss the topic in the very details. Nevertheless, we hope that reader will be inspired by the topic and pay attention to this area or bear macro-economy on mind when solving the public health development strategy as one of the determinants and limits playing important role during implementation of interventions.

References

1. DAVID, E., BLOOM, D. C. a. J. S. The Effect of Health on Economic Growth: A Production Function Approach. World Development, 2004. 32(1): p. 1 – 13.
2. EUROSTAT, EUROPEAN SYSTEM OF ACCOUNTS ESA 1995. 1995: Brussels.
3. EC, I., OECD, UN, WB, System of national Accounts 2008. 2008.
4. Michal Baránik, J. H., Fundamentals of Macro-Economy. 2002, Trenčín: Vydavateľ Iris.
5. Lisý, J. et al. New Economy 2007: IURA EDITION.
6. Oldřich Hájek, P. G., Boris Popesko, Šárka Hrabinová, Czech Republic vs. EU-27: Economic Level, Health Care and Population Health. Cent Eur J Public Health, 2012. 20(3): p. 167–173.
7. Muntaner, C, C. H., Benach, J., Ng, E., Hierarchical cluster analysis of labor market regulations and population health: taxonomy of low – and middle-income countries. BMC Public Health, 2012. 12(286): p. 1 – 15.
8. Roy Carr-Hill, E. C., What explains the distribution of doctors and nurses in different countries, and does it matter for health outcomes? Journal of Advanced Nursing, 2013. published ahead of print: p. 1 – 13.
9. Health protection: Accident prevention and injury control. Public Health Rep., 1983. Suppl: p. 80 – 93.
10. Source: Statistical Office of the Slovak Republic. Regional Statistics Database RegDat. 2013 [cited 2013 22.5.]; Available from: http://px-web.statistics.sk/PXWebSlovak/.
11. SOURCE: STATISTICAL OFFICE. System of demographic statistics and foreign migration statistics in Slovakia in 2011 [cited 2011 25.2.]; Available from: http://portal.statistics.sk/.
12. OECD, E., Methodological Manual on Purchasing Power Parities. 2005: OECD.
13. WHO, Part Iii Global health indicators, in World Health Statistics 2013. 2013, World Health Organization: Geneva.
14. WHO. Global Health Expenditure Database, GHED. 2013 [cited 2013 9.7.]; Available from: http://apps.who.int/nha/database/DataExplorerRegime.aspx.
15. WHO. WHO Indicator and Measurement Registry (IMR). 2013 [cited 2013 9.7.]; Available from: http://apps.who.int/gho/indicatorregistry/App_Main/indicator_registry.aspx.
16. UNDP, Human Development Report 1990. 1990, United Nations Development Program (UNDP): New York, Oxford, Oxford University Press.
17. CPHR, C.p.h.r. National Report of Human Development – Slovak Republic 2001 – 2002. 2002 [ci-ted 2013 17.06.]; Available from: http://www.cphr.sk/undp2002.htm.
18. OECD. The Survey of Adult Skills (PIaac). 2013 [cited 2013 8.7.2013]; Available from: http:// www.oecd.org/site/piaac/surveyofadultskills.htm.
19. UNDP, Human Development Report 2013: The Rise of the South: Human Progress in a Diverse World. 2013: UNDP

Chapter 12

Mathematical models and public health

Chapter objectives

In the previous chapters we discussed healthcare statistics and its methods used in the public health study. We know that health and illness doesn´t result from a single factor influence but co-influence of a complex of factors. To specify these statements, we should briefly think about factors affecting mortality rate development in case of cardiovascular diseases. All of us will probably think about unacceptable lifestyle at first place. The studies point out multiple increase of heart attack risk of a man who smokes, doesn´t have enough physical activity, has increased cholesterol level in blood as a result of poor nutrition characterized with insufficient intake of fruit and vegetable, among other nutrients. [1] Thus, we could consider that these risk factors explain mortality rate development in case of cardiovascular diseases in time (the higher risk factors prevalence, the higher mortality rate). In such considerations we shouldn´t forget further factors that could play an important role here. Fundamentals of the theory of determinants say that environment, healthcare services quality and biological/ genetic factors influence health and illness, along with the lifestyle. It is apparent that healthcare services effect should be included in our considerations. How many people pass cardiologic surgery or interventions? How many people are treated with pharmaceuticals that postpone death that would have been caused by cardiovascular diseases much earlier if the pills hadn´t been administered? We should also think about the effects of social – economic health determinants. Our considerations indicate that the problem examined by us is actually a consequence of a complex of determinants. We could state also another example of demanding questions: Politicians are interested in the expenditures necessary to put aside for healthcare in the following period.

A hospital director would like to know how many patients approximately will have to be hospitalized during infection epidemics. If a public healthcare officer wants to enforce implementation of a preventive measure, he should possess a proof of evidence of the measure efficiency. In other words, he should provide information to people authorized to decide on what will happen if the intervention is/ isn´t implemented and what would its benefits be alike. Epidemiological studies are a traditionally used identification method of disease occurrence cause and efficiency of related prevention/ therapy methods. Epidemiological studies often bear unbeatable obstacles resulting from ethical, financial or other implementation restrictions. Some of them could be surpassed using mathematical modeling methodology.

This chapter is aimed at providing a reader with basic information about mathematical models on theoretical level and their application to the public healthcare system. Studying through the chapter, a reader will be able to understand the principles of biological systems modeling, the mathematical model structure and to describe many recognized mathematical models.

Definitions of mathematical model

Brief history

In 1766, Swiss mathematician and physicist Daniel Bernoulli (1700 – 1782) published an article where he tried to find out the progress of life expectancy in France after established inoculation against varicella; upon request of the mathematician Maupertuis. At the time, there were discrepancies between the varicella vaccination defendants and opponents. It was discussed whether mandatory vaccination should be established despite of having sometimes fatal consequences. Thus, it was about the way of comparing potential vaccination benefits on long-term basis vs its actual risks. With calculations, Bernoulli found out that varicella vaccination is beneficial only if the fatality risk is below 11 % and vaccination increases life expectancy by more than 3 years.

Bernoulli is considered a creator of the first epidemiological model applicable to infectious diseases. [2, 3] Mathematical modeling was fully developed in the computer era when complicated calculations could be run with high accuracy and speed.

Mathematical models

Critchley and Capewell define the model as „simplification of reality, which varies from simple, descriptive tools (e.g. house drawing) towards the system of mathematical procedures that are able to explain previous development of disease occurrence or to predict future events, for example epidemics“ [4]. Weinstein defines mathematical model as a „logical, mathematical system enabling to enter facts and values therein and joining the input data in the results that people authorized to decide on healthcare are interested in". [5]

There is another model definition saying that the model is "simplification of reality, which is aimed at helping answer specific questions ". [6]

We could come to conclusion through the synthesis of both definitions: modeling is an approach using mathematical language that can help at research of complex healthcare issues, provided it is used adequately and both authors and users of the model outcomes are aware of its limits. While the model isn´t capable of „providing the only correct answer", its most important task is to turn people authorized to decide to correct direction, namely to offer possible solution. [6]

Regress models

To avoid impression at a reader that the model should be extremely complicated, we should mention models used relatively often in the public health studies. Models of regress analysis represent probably the most recognized ones and amongst them we should state linear and logistic regression. Both types of regression are described in detail and discussed in the publication about bio-statistics [7] and further publications dedicated to statistics. Detail description of the models exceeds the chapter goals. We should just repeat that linear regression describes relation between two or more variables. Contrary to correlation, linear regression allows for moving a step ahead and find out how a variable marked "dependant" (Y), depends on the other one (or more) independent variables (X). To illustrate it, we could state an example of relation between mortality caused by lung cancer and the count of sold cigarettes. We could expect positive relation (the higher count of sold cigarettes the higher mortality for lung cancer). Such described relation could be then moved ahead and we should ask what the mortality for lung cancer will be if e.g. 10 million of cigarettes are sold. This is the regression principle – we create a model from our data and it is used for prediction of dependant variable relation (mortality) in case of changed independent variable (number of sold cigarettes). In case of linear regression, it is a linear model, which means we will plot a line through our scatterplot and place it so that the distance between the line and each point on the scatterplot shall be as short as possible. How can we do it? It is apparent that we cannot apply our intuition since it is often subjective. Therefore we have to use mathematical model called "least squares method". In this way, we can get a line that best covers the data. Statistical programs and program environment {R} are able to calculate characteristics of such line, namely: intersect (*a*), i.e. point in which the line intersects the axis y at the moment of X variable having value 0, and the line slope (*b*), presenting the value of variable Y at the change of X variable by one.

Model of linear regression is characterized by formula:

$$Y_i = a + bX_i + \epsilon_i \,,$$

where Y_i je refers to dependant variable value, a is intersect, b is regression coefficient characterizing the line slope, X is independent variable value and ϵ_i is error representing the difference between the value predicted by the line and actually observed value (the formula is sometimes presented without error but we should take in account that it is error that points out to imperfections of the derived relation; in fact it is a model).

Thus, if we have data and we calculated the values of regression coefficients (a) and (b), we can replace the independent variable in the formula with any number and predict the value of dependant variable. We created a model of simple linear regression. As we said, there are very few effects in the world that result from only one factor of influence. Therefore, the model of simple linear regression refers to huge simplification of reality as well. It would be better to extend the model with further explaining variables (X_1, X_2, ...X_i) and to create model of multiple linear regression.

Linear regression is conditioned with dependant variable being a quantitative continuous variable. But what if we are interested in dependant variable that is categorical (e.g. of type yes - no)? Let´s show an example. Currently, cardiovascular diseases are one of the most frequent diseases in the developed countries and it is important to know suitable time for a physician when exactly the patient´s therapy should start. The physician should consider the risk of cardiovascular disease likelihood at the patient (e.g. heart ischemia) as well as incurred cost. The disease likelihood is an essential question here (will or will he not fall ill?). How could the physician identify this likelihood: Currently, there are tools based on big epidemiological studies. If we have a representative set of people suffering heart ischemia and those healthy where we monitor blood pressure, cholesterol concentration, smoker/ non-smoker status, gender and other characteristics, we will be able to determine factors that caused heart ischemia at particular patients. Using logistic regression, we can find out that all above factors could cause heart ischemia (we mostly monitor the factors´ effect in logistic regression through odds ratios). This procedure allows for prediction of likelihood of an individual not included in our set to fall ill for heart ischemia. If we measure his blood pressure, actual cholesterol concentration, ask him if he smokes, know his gender, and we enter all the data in the logical regression model, we can calculate the likelihood of heart ischemia occurrence at particular individual. Thus, we see that logistic regression serves as a mathematical model.

Model categories

Models are divided in the quantity and quality ones. Quantity models aren´t structured to deliver information in the form of numbers; instead they provide us with overview of the model function and the reason of its existence. Health belief model is an example of quality models, used for description of human behavior mainly when planning interventions. Quantity models can be called mathematical models since they use language and tools of mathematics to describe the system behavior.

Speaking of time, we can divide them in static and dynamical. Static models provide intersection of the situation in particular time point (or two points). As such, they don´t consider time effect on the monitored variables. Anyway, their results are often sufficient because of simple structure. Contrary to the static models, the dynamical ones allow for change of variables in time. Thanks to the time component, dynamical models allow for simulation of monitored effect evolution. However, their structure is more complicated. [6]

Other models structure divides them in discrete and continuous. Discrete models enable change of status (e.g. smoker – former smoker) only in certain time sections (e.g. once per a year). Continuous models enable the changes at any time.

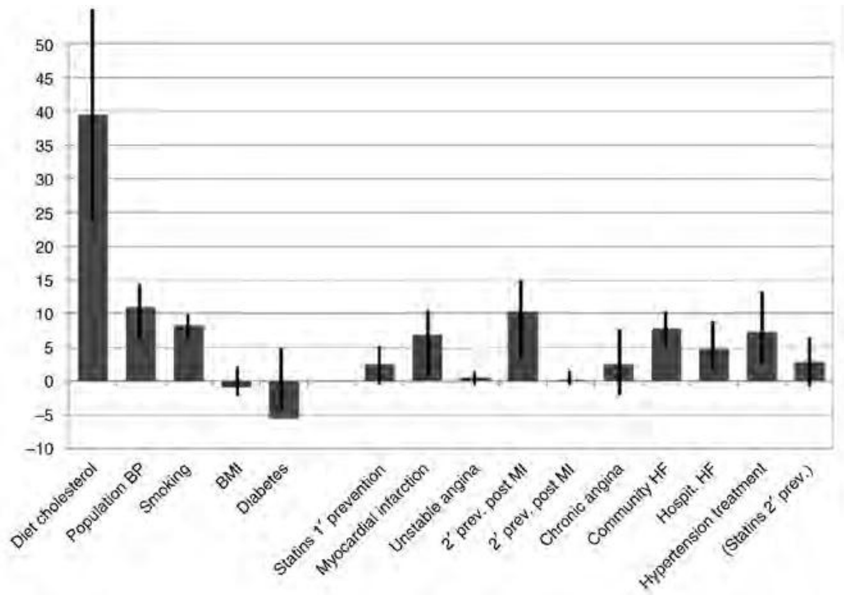
Models can be further divided in micro and macro simulations. Macro simulations distinguish for example social – economic groups and groups divided according to gender or income. Contrary to macro simulations working with proportions in the group, micro models work with individual characteristics. [8]

We distinguish between deterministic and stochastic models according to the way of change occurrence. Stochastic models work with random events and behavior. Deterministic models work with fixed and predictable events and behavior. [6] In other words, deterministic model releases always the same results upon entered same values while stochastic model delivers different results based on random entries.

Process of model creation

In general, model creation process includes a few steps. At first, we should define the problem in question. Then we choose a suitable model type; specify its basic structure and relations between input data and model outputs (mathematically expressed). In this moment, input data become of key importance since data collection, processing and analysis represents the second step. We should remind already mentioned slogan „Garbage in, garbage out“, suggesting in this case that a model is only of such quality as its input data. The data refer to critical aspect of any model creation; they are often not available and additional studies are required to obtain the data or one should rely upon expert and explicit estimations.

Model validation is the next step, i.e. verification if the model works as we want it to. Model works properly if its outputs are in compliance with historical and empirical findings (for example, if we want the model to explain postponed death; let´s say we want to explain 1000 of such deaths. The model is validated if most of them will be explained.). Since we enter data from various sources in the models, often of secondary nature (i.e. not collected for purposes of the model), we introduce errors in the model. This requires sensitivity analysis examining the difference in the model results in case of entered various values of the same parameter. Example of the sensitivity analysis can be seen on picture No. 1. It is a sensitivity analysis from IMPACT model, where we can see the effect of applied therapeutic methods and risk factors on decreased mortality rate in case of heart ischemia in the Czech Republic during period of years 1985 – 2007. We can see that blood cholesterol has the strongest influence on decreased mortality rate. Vertical line doesn´t represent confidence intervals but results of the sensitivity analysis. Entering various values (top and bottom extreme values) of average total cholesterol concentration in the population, we can see that cholesterol is the strongest factor. Blood pressure and smoking didn´t keep their influence rate in the sensitivity analysis.[9] If the model generates approximately the same results in various conditions, we call it "robust". Model improvement represents the last stage since the model creation isn´t only the process of seeking answers to questions asked but also a learning process. [6, 8]



Picture No. 1 Sensitivity analysis in the model IMPACT. Taken from [9]

Use of mathematical models in public healthcare system

Mathematical modeling is widely used in the area of healthcare and public healthcare study. It includes:

• Estimation of the future healthcare needs, healthcare provision optimizing • reduction of waiting times and healthcare unavailability, • examination of demographical effects on the population health condition, • understanding and monitoring of infectious diseases spread (for example influenza), • prediction of future diseases occurrence in the population (for example diabetes mellitus, obesity), • explanation of incidence of or mortality for various diseases, • explanation of gained life years, saved funds as a result of applied intervention or, expression of adverse effects of the planned action (e.g. policy).

Table No. 1 Use of mathematical models in healthcare system

It is apparent that mathematical models represent a very useful decision-making tool. Healthcare cost has been constantly increasing. Accordingly, taking in account also the pressure applied on quality of provided healthcare services, it is necessary more than ever that the policy authors and persons authorized to decide make decisions based on the knowledge instead of the random ones. Questions appearing in the area of healthcare are so complex that it isn´t acceptable anymore to reply upon intuition or random answers. Mathematical modeling has become an important tool in search for solid, defendable and proof-based answers to such complex questions. [6]

In the following text we will describe a few models used in practice in order to demonstrate statements in the table No. 1.

Method Monte Carlo

Method Monte Carlo[38] represents a system of mathematical processes using random numbers and problem solving probability. The method is widely used in many areas from economy through physics up to public healthcare system study. It is a stochastic method used mainly if deterministic approaches are insufficient to describe events strongly affected by random inputs. The Monte Carlo method principle is easy: we seek a median of the parameter as a result if random event.

[38] The name was derived from well known city Monte Carlo, known for casino and roulette

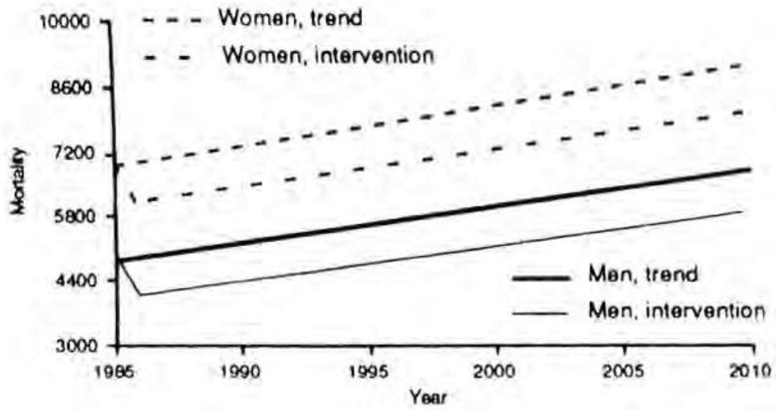
We know from the theory of statistics that we have to calculate median from many input data. The Monte Carlo method is about creation of a model that shall be repeated hundreds or thousands times and the inputs are generated randomly within set forth real range. Each repetition brings required output as a consequence of entered (random) inputs. As soon as we have enough repetitions, we can calculate median as well as standard deviation, and we are close to the real output that takes in account uncertainties affecting it in real life. This method is widely used in the area of public health. Monte Carlo method was applied for example to description of transmission of meticiline resistant Staphylococcus aureus (MRSA) among patients in hospital ward, depending on a healthcare assistant´s hand-hygiene habits, [10] or during examination of air pollution influence on illness and mortality rate. [11]

PREVENT

Model **PREVENT** was developed in the 80s of the 20th century. [12, 13] This macro-simulation estimates potential benefits (in the form of averted death and gained life years) resulting from primary prevention – reduction of chronic diseases risk factors prevalence in the population. Model application allows us to compare how mortality rate would evolve for chosen disease group in case of reached reduction of risk factors prevalence through adopted vs not adopted preventive measures. Model advantages can be summarized as follows: it takes in account demographical changes and enables long-term simulations, includes classical risk factors and various groups of chronic diseases, it is user-friendly, updated method versions include calculation of cost and estimation of illness rate in case of particular diseases. Adverse features result from the model architecture since addition of a new risk factor is very complicated. [4]. Model PREVENT has been still widely used. For example, lately it was used in the project EURocadet,[39] which estimated future incidence of carcinoma diseases in case of implemented interventions. Furthermore, model PREVENT was used in the project EPIderM,[40] for estimation of potential effect of interventions focused on reduction of UV radiation on skin cancer incidence in five European countries. [14] Currently, the model was used to estimate the project effectiveness rate in Northern Ireland[41] focused on cultivation of environment, building of a park and bike routes in the city, in order to increase the citizens´ physical activity. [15]

---

39 http://www.eurocadet.org/objective-3.html#note
40 http://www.epiderm-network.eu/home.html
41 http://www.communitygreenway.co.uk/project/about-connswater-community-greenway



Picture No. 2 Number of deaths caused by cerebrovascular embolism that would occur if intervention was/ wasn´t made, aimed at fighting with hypertension according to Prevent model. Taken from [13].

Model IMPACT

Deterministic **model IMPACT**, explaining the causes of heart ischemia mortality rate rise or drop, is based on the Excel program functions. The model considers occurrence of heart ischemia risk factors in the population and application of proof-based heart ischemia therapeutic methods; quantifying the effect which the therapeutic methods contribute to mortality rate rise or drop during specified period. Based on the heart ischemia caused deaths and the population size in the start and end year of the monitoring, number of death shall be calculated that should occur in the end year of the monitoring if the mortality rate reported in the start year of monitoring was unchanged. However, such number didn´t occur since the mortality rate decreased and actually observed number of deaths was derived from the expected number. Result of the difference refers to „deaths prevented or postponed (DPPs)“. Such DPPs number should be explained by the model. As we stated above, the model works with two groups of explaining factors: prevalence of risk factors and application of therapeutic methods. As for heart ischemia risk factors, the model works with smoking prevalence, with average total cholesterol concentration, with body mass index (BMI), systolic blood pressure, prevalence of physical inactivity and diabetes. Furthermore, the model works with nine patient categories (e.g. patients hospitalized with acute heart attack, those living with chronic heart failure, etc.). [12] Model IMPACT was developed in 1996 [16] and its basic version has been built in 19 countries since then. Results of the model point out to the fact that heart ischemia – caused mortality rate drop has been mostly

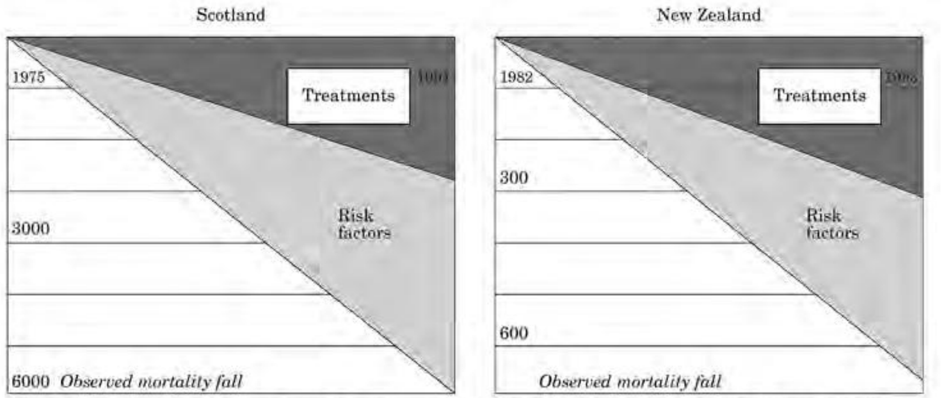


Figure 3 **Declines in coronary heart disease mortality in Scotland 1975–94 and in New Zealand 1982–93: Estimated contribution of medical treatments and risk factor reductions[38,39].**

Scotland		New Zealand	
Total treatments 40%	Total risk factors 60%	Total treatments 46%	Total risk factors 54%
AMI treatments 10%	Smoking 36%	AMI treatments 12%	Smoking 30%
Angina (revascularization) 2%	Cholesterol 6%	Angina (revascularization) 4%	Cholesterol 12%
Angina (aspirin) 2%	Population BP (15%)*	Angina (aspirin) 3%	Population BP (11%)*
Heart failure 8%	Deprivation 3%	Heart failure 7%	Deprivation — N/A
Hypertension treatment 9%	Other factors 9%	Hypertension treatment 7%	Other factors 4%
Secondary prevention 8%		Secondary prevention 12%	

*Population blood pressure includes the effect of treatment of individuals for hypertension

Picture No. 3 IMPACT model results in Scotland and New Zealand. Taken from [4]

caused by the changes at prevalence of modifiable risk factors and application of proof-based therapeutic methods. 23 – 47 % of observed mortality rate decrease can be attributed to these facts. The results are extremely important for planning of cardiovascular diseases fighting strategies. Example of IMPACT model results is stated on picture No. 3. We can see that heart ischemia caused mortality rate decreased in both countries during the monitored period, as reflected in DPPs. 60 % of the decrease in Scotland can be attributed to decreased prevalence in risk factors and approx. 40 % to therapeutic methods. As for New Zealand, 54 % DPPs can be explained with risk factors prevalence and 46 % to therapeutic methods. [4]

PROS and CONS of mathematical models use

Sometimes the public healthcare sector faces complicated issues and use of classical epidemiological studies would be very demanding or even impossible (remind the Bernoulli problem). Modeling enables bridging the limits of epidemiological studies where the following facts can be included in [4]:

- Demographical limits – only a few epidemiological studies include adequate proportion of women among the participants, ethnic groups or elderly people. Such groups could be included in the model through explicit estimation of risk in these groups.
- Methodological limits – to be strong enough, epidemiological studies are time and cost demanding with ever threatening risk of bias. For example, results of randomized control trials are hard to apply in general to the whole population. Such fact results from strict entry criterions of the trial participants. The models can help evaluate the importance of such biases, for example through transfer of the trial results to real, general population.
- Data specific limits – models can help at research of potential importance of factors where proofs of evidence aren´t sufficient or doesn´t exist at all.
- Limits resulting from complexity of effects – models can help compare potential benefits of various approaches and enable policy authors and persons authorized to make decisions to identify the areas of research priority.

Moreover, the ability of models to combine data from various sources and disciplines and thereby contribute to creation of "complex view" of a monitored effect is considered a strong point of the models. Models should be used also in the situation where making many experiments to identify the value of parameter in question is not feasible (for example Monte Carlo method). Modeling disadvantage lays in the fact that its results strongly depend on the input data quality, as well as preconditions that should be made in case of unavailable input data. There is also another disadvantage – the model outcomes recipients – managers, policy authors and persons authorized to make decisions – could have problems with understanding of mathematical models functioning and thus prone to untrust the results.

Summary

In this chapter we discussed mathematical models and their use on the public health research. We defined the model and dealt with various aspects of model classification. We said that the mathematical models can vary from simpler recognized models, e.g. regression analyses, up to complicated models whose creation requires deep knowledge of the area of solved problem, as well as knowledge of advanced mathematical apparatus. We found out that the models are an efficient tool of healthcare management and their use allows for decision making based on proofs of evidence instead of random decision. We presented a few well-known models and the Monte Carlo method. Comparing the models PROS and CONS, we came to conclusion that the models could fill the gaps in knowledge and combine data from various sources in a whole.

Compared to epidemiological studies (mainly costly cohort studies), modeling is a cheaper and faster alternative. However, when used, we should take in account its limitations.

References

1. Yusuf, S. et al. *Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study.* Lancet, 2004. **364**(9438): p. 937 – 952.
2. Dietz, K. and Heesterbeek, J. A. P. *Daniel Bernoulli's epidemiological model revisited.* Mathematical Biosciences, 2002. **180**: p. 1 – 21.
3. Bacaër, N. *Daniel Bernoulli, d'Alembert and the inoculation of smallpox (1760)*, in *A Short History of Mathematical Population Dynamics*. 2011, Springer London. p. 21 – 30.
4. Critchley, J. A. and Capewell , S. *Why model coronary heart disease?* European Heart Journal, 2002. **23**(2): p. 110 – 116.
5. Weinstein, M. C. et al. *Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices-modeling studies.* Value in Health, 2003. **6**(1): p. 9 – 17.
6. The Complex Systems Modeling Group, *Modeling in Healthcare.* 2010, Providence,Rhode Is-land: American Mathematical Society.
7. Rusnák, M., Rusnáková, V., and Majdan, M., *Bio-statistics for public healthcare system students*. 2010, Trnava: Typi Universitatis Tyrnaviensis – Publishing agency of Trnava University.
8. Levy, D. T. et al. *Simulation models of obesity: a review of the literature and implications for research and policy.* Obesity Reviews, 2011. **12**(5): p. 378 – 394.
9. Bruthans, J. et al. *Explaining the decline in coronary heart disease mortality in the Czech Republic between 1985 and 2007.* European Journal Of Preventive Cardiology, 2012.
10. Beggs, C., Shepherd, S. and Kerr, K. *How does healthcare worker hand hygiene behaviour impact upon the transmission of MRSA between patients?: an analysis using a Monte Carlo model.* BMC In-fectious Diseases, 2009. **9**(1): p. 64.
11. MacDonald Gibson, J. et al. *Deaths and Medical Visits Attributable to Environmental Pollution in the United Arab Emirates.* Plos One, 2013. **8**(3): p. e57536.
12. Unal, B., Capewell, S. and Critchley, J. *Coronary Heart Disease Policy Models: a systematic review* BMC Public Health, 2006. **6**(1): p. 213.
13. Gunning-Schepers, L. J., Barendregt, J. J. M. and van der Maas, P. J. *PREVENT, a Model to Estimate the Health Benefits of Prevention*, in *Models of Non-communicable Diseases. Health status and Health Service Requirements*, W. Morgenstern, et al., Editors. 1992, Springer – Verlag: Berlin. p. 55 – 69.
14. de Vries, E. et al. *Potential impact of interventions resulting in reduced exposure to ultraviolet (UV) radiation (UVA and UVB) on skin cancer incidence in four European countries, 2010 – 2050.* British Journal of Dermatology, 2012. **167**: p. 53 – 62.
15. Dallat, M. A. T. et al. *Urban greenways have the potential to increase physical activity levels cost-effectively.* The European Journal of Public Health, 2013.
16. Capewell, S., Morrison, C. E. and McMurray, J. J. *Contribution of modern cardiovascular treatment and risk factor changes to the decline in coronary heart disease mortality in Scotland between 1975 and 1994.* Heart, 1999. **81**(4): p. 380 – 386.

Chapter 13

Conclusion

The monograph contains statements and opinions of two recognized domestic authors and a promising young specialist in the area of public healthcare sector. The authors also confronted their knowledge and results of work with those of foreign colleagues, documented through cited literature. All three authors professionally deal with the health statistics and possibilities of use of modern statistical and SW tools application methods and procedures. Based on the professional statements and opinions, a reader can get more profound and complex view on the health statistics as the basis of creation of strategies and policies in the area of public health.

The monograph benefits could include preparation of sufficient amount of information and practical examples mainly related to methods of data acquisition, processing and interpretation.

Another benefit refers to generalization and transparency of particular principles of identification of health services quantity signs in both healthcare and public health context in its new paradigm. Information, examples and facts contained in the monograph can be applied to scientific research of facts and their mutual correlations, as well as in the educational process.

Index

Statistics of Health

Martin Rusnák, Viera Rusnáková, Marek Psota

Add titles

Editor-in-chief: PhDr. Jozef Molitor
Graphical design and a composition Jana Janíková
Book jacket draft ????????

For the Faculty of Healthcare and Social Work Trnava University
published by agency Typi Universitatis Tyrnaviensis,
common office of the Trnava University in Trnava
and SCIENCE, publishing agency o the Slovak Academy of Sciences as the131st publication. Printed by VEDA, SAV publishing agency.

ISBN 978-80-8082-732-8