More information available at:
www.iba.muni.cz/summer-school2013

Publication was supported by the ESF project no. CZ.1.07/2.2.00/28.0043
"Interdisciplinary development of Computational Biology study programme"
and national budget of the Czech Republic.

INVESTMENTS IN EDUCATION DEVELOPMENT

ISBN 978-80-7204-756-7

SVRATKA 2013

Proceedings of the 9th Summer School on Computational Biology

Institute of Biostatistics and Analyses
Masaryk University

Proceedings of the 9th Summer School on Computational Biology

Stochastic Modelling in Epidemiology

10–13 September 2013
Svratka, Czech Republic

Editors:
Tomáš Pavlík
Ondřej Májek

INVESTMENTS IN EDUCATION DEVELOPMENT

Institute of Biostatistics and Analyses
Masaryk University

Proceedings of the 9th Summer School on Computational Biology

Stochastic Modelling in Epidemiology

10–13 September 2013
Svratka, Czech Republic

Editors:
Tomáš Pavlík
Ondřej Májek

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitivness
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTMENTS IN EDUCATION DEVELOPMENT

ISBN 978-80-210-6305-1

Contents

Preface

The 9th year of the Summer School on Computational Biology continues in a yearly tradition of informal summer schools focused on interesting aspects of biology, health care research, and biomedicine. This year's theme is "Stochastic Modelling in Epidemiology". Bearing in mind the broad definition of epidemiology: "Epidemiology is about the understanding of disease development and the methods used to uncover the etiology, progression, and treatment of the disease", we can consider the scope of epidemiology in public health being as old as mankind itself. However, it definitely does not mean that epidemiology is out of fashion. In fact, the opposite is true. Epidemiology has an indisputable role in clinical research, where the methods of epidemiology still contribute to more and more detailed understanding of the processes associated with different diseases. Nowadays, modern epidemiology cannot be imagined without statistical methods that help us uncover the hidden associations between factors and diseases under study. Stochastic models belong among the main procedures used in this way. These models can help us in quantifying factor effect, adjusting for confounding variables, and studying complex correlation structures. Therefore, stochastic models in epidemiology represent an up-to-date issue that we hope will be of interest to all participants of the 9th year of the Summer School in Computational Biology.

We greatly acknowledge financial support by the Ministry of Education, Youth and Sports of the Czech Republic; project CZ.1.07/2.2.00/28.0043 "Interdisciplinary Development of Computational Biology Study Programme" and national budget of the Czech Republic.

On behalf of the programme and organizing committee,

Brno, August 12, 2013

Ondřej Májek

Tomáš Pavlík

Stochastic Modelling in Epidemiology

Lectures

Basic concepts in epidemiology

Ondřej Májek

Institute of Biostatistics and Analyses, Masaryk University, Brno;
e-mail: majek@iba.muni.cz

Abstract

Epidemiology studies the distribution and determinants of disease in human populations. In this lesson, basic concepts of epidemiology will be explained. After a short introduction, measures of disease occurrence will be defined in Chapter 2. Chapter 3 will introduce measuring changes in disease frequency as a result of exposure of individuals to a particular risk or protective factor. In epidemiology, several study types with various advantages and disadvantages may be used to measure such effects. These study types will be introduced in Chapter 4, along with sources of error in such studies. Validity of a study may be also compromised due to confounding, a phenomenon of key importance in the observational epidemiology. Confounding variables will be described in Chapter 5. Chapter 6 describes basic methods of adjusting for confounding variables. Basic methods of data analysis in epidemiology will be explained in this lesson; however, detailed description of more advanced statistical methods will be given in the following lessons.

Key words

Epidemiology, occurrence measures, effect measures, study types, confounding.

1. Introduction

Epidemiology can be defined as the study of the distribution and determinants of disease frequency in human populations (Rothman et al., 2008). This definition is reflected in the basic concepts described in this lesson: firstly, measuring occurrence of a disease in a population, followed by measuring effect of the *exposures* (when looking for the cause of a disease, such as smoking, alcohol, polluted environment, etc.) or *interventions* (e.g., new medicine, medical procedure, vaccination, screening programme etc.) on risk of the disease. This task is not an easy one, as many associations between exposures and diseases may be spurious, not providing an opportunity to intervene and improve the health of individuals. Finding the causal relationship between a disease and exposure includes both proper design of an epidemiological study, which needs to be appropriate for the presented problem, and the application of proper statistical methods for the analysis of collected data. Although the classical epidemiology started with communicable diseases (one of the most notable early achievements is the work of John Snow, who elucidated the association between cholera and the source of drinking water in the 19th century), epidemiologists have also paid a lot of attention to chronic diseases (e.g., cardiovascular, cancer, diabetes, etc.) and their causes (Saracci, 2010).

2. Measures of disease frequency

The objective of epidemiology is often to quantify the effect of a potential cause on the occurrence of disease. Therefore, we must be able to measure the frequency of disease occurrence. The first step in calculating disease frequency is to specify the population under study, usually including only persons susceptible to a given disease (for example, omitting females when prostate cancer is the subject of interest). *The population at risk* can be defined by demographic, geographic or environmental factors. In this chapter, the term *disease case* also relates to injuries, different physiological changes under study etc. Definitions of basic disease frequency measures are presented below (dos Santos Silva, 1999; Bonita et al., 2006).

Prevalence of the disease is calculated as follows:

$$Prevalence = \frac{\text{No. of existing cases in the population at risk at one point in time}}{\text{No. of people in the population at risk at the same point in time}}$$

In practice, data on the population at risk are often approximated by the total population in the study area. The above-mentioned definition relates to the so-called *point prevalence*, as it refers to a single particular point in time. Accordingly, we may refer to an entire specified period of time. In that case, the measure is called *period prevalence.*

Incidence risk of the disease is calculated as follows:

$$Incidence\ risk = \frac{\text{No. of new cases of disease arising in the population at risk over a given period of time}}{\text{No. of disease-free people in that population at the beginning of that time period}}$$

Calculation of incidence risk assumes that the entire population at risk at the beginning of the study period was followed up during the whole period of time. To account for varying lengths of follow-up, the denominator can be recalculated to represent the sum of individual times that a particular person was at risk of becoming a case. This is called *person-time at risk.* The resulting occurrence measure is usually called *incidence rate* (as opposed to the incidence risk, which is a *proportion*). Thus, the *incidence rate* of the disease is calculated as follows:

$$Incidence\ rate = \frac{\text{No. of new cases of disease arising in the population at risk over a given period of time}}{\text{Total person-time at risk during that period}}$$

Again, in practice, the denominator is often calculated approximately by multiplying the average size of the study population by the length of the study period. The difference between risk and rate is depicted in Figure 1.

Provided that the prevalence is low and does not change significantly over time, it can be calculated as a product of incidence and average duration of the disease (Bonita et al., 2006).

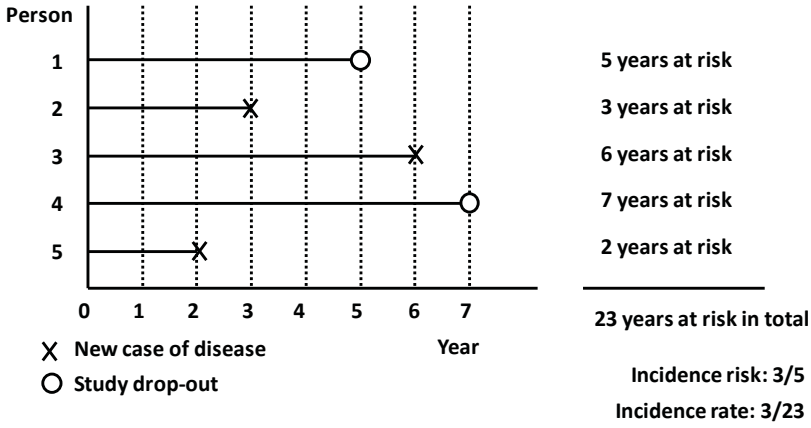


Figure 1. The outline of computation of incidence risk and incidence rate.

Mortality is an indicator similar to incidence, taking death as a specific case. The severity of a disease, represented by the proportion of persons who die within a specified time, is shown by *case fatality*.

$$\textit{Mortality rate} = \frac{\text{No. of deaths occurring in the population at risk over a given period of time}}{\text{Total person-time at risk during that period}}$$

$$\textit{Case fatality} = \frac{\text{No. of deaths from diagnosed cases in a given period}}{\text{No. of diagnosed cases of the disease in the same period}}$$

3. Measures of effect

An *effect* of a factor is the change in a population characteristic (such as incidence or mortality) that is caused by the factor considering one of its levels versus another (Rothman et al., 2008). Characteristics in epidemiology with potential effects on disease frequency are usually called *exposures* (e.g. behaviour, treatment or other intervention, trait, exposure in the usual sense, or some other disease). The effect measures in epidemiology could be either *relative* (represented by the *ratio* of two disease incidences – exposed group vs. unexposed group) or *absolute* (represented by the *difference* of the two disease incidences).

3.1 Relative measures of exposure effect

Relative measures are used to ascertain the strength of association between the exposure and the outcome of interest, i.e., how much is the exposed group more likely to develop the disease compared to the unexposed group. Computation of basic effect measures is particularly simple from a 2×2 contingency table of binary outcome against a binary exposure status (Table 1). Statistical significance of the association between the outcome and the exposure could be tested using either the chi-square test or the Fisher's exact test

(Woodward, 1999). We can also use this table to compute the *risk ratio*, *odds ratio* and related effect measures (dos Santos Silva, 1999).

Table 1. General representation of the study results.

		Exposed group	**Unexposed group**	**Total**
Outcome present	Yes	*a*	*b*	*a+b*
	No	*c*	*d*	*c+d*
	Total	*a+c*	*b+d*	*n*

Generally, the risk ratio (*RR*) is calculated as follows:

$$RR = \frac{\text{Risk of disease in the exposed group } (R_1)}{\text{Risk of disease in the unexposed group } (R_0)}$$

Using the notation and data introduced in Table 1, we can compute the *RR* as follows:

$$RR = \frac{R_1}{R_0} = \frac{a/(a+c)}{b/(b+d)}$$

The distribution of the sample risk ratio is skewed and we will therefore use a log transformation to ensure approximate normality. The following formulas show a standard error and 95% confidence interval (CI) for sample log*RR*. To construct CI for *RR* itself, we will exponentiate the two CI limits (Woodward, 1999).

$$SE(\log RR) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$$

$$\log RR - 1.96 \times SE(\log RR) \quad \text{to} \quad \log RR + 1.96 \times SE(\log RR)$$

The definition of risk just used is identical to probability. In epidemiology, it is often useful to use one more specification of chance called the *odds*. The probability is computed as the number of times at which a specified outcome is present, related to the total size of the group. On the other hand, odds are calculated as the number of times at which a specified outcome is present, related to a number of times at which the same outcome is not present. Similarly to risks, we can compute odds ratio (*OR*) relating the occurrence of events in two groups:

$$OR = \frac{\text{Odds of disease in the exposed group}}{\text{Odds of disease in the unexposed group}} = \frac{\dfrac{R_1}{1-R_1}}{\dfrac{R_0}{1-R_0}}$$

Using the notation and data introduced in Table 1, we can compute the *OR* as follows:

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Similarly to the relative risk, we will approximate the log transformed *OR* by normal distribution:

$$SE(\log OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\log OR - 1.96 \times SE(\log OR) \quad \text{to} \quad \log OR + 1.96 \times SE(\log OR)$$

Risk is usually the preferred measure because it is easily understood by most people. On the other hand, *OR* is used very often in epidemiology. In case-control studies (see below), we can only estimate *OR*. Moreover, *OR* will be a good approximation of *RR* whenever the disease in question is rare. It is, however, necessary to acknowledge that these measures may differ substantially for common diseases.

***

Example 1. The following table shows results of Pooling Project that studied risk factors for coronary heart disease in men (Woodward, 1999):

		Smoker at entry		
		Yes	No	Total
Coronary event during follow-up	Yes	166	50	216
	No	1176	513	1689
	Total	1342	563	1905

The risk ratio is $$RR = \frac{166/(166+1176)}{50/(50+513)} = 1.393$$

The standard error of the log(*RR*) is

$$SE(\log RR) = \sqrt{\frac{1}{166} - \frac{1}{166+1176} + \frac{1}{50} - \frac{1}{50+513}} = 0.1533$$

The lower and upper limit of the 95% CI of log(*RR*)

$$L_{\log RR} = \log 1.393 - 1.96 \times 0.1533 = 0.031 \quad L_{RR} = \exp(0.031) = 1.031$$

$$U_{\log RR} = \log 1.393 + 1.96 \times 0.1533 = 0.632 \quad U_{RR} = \exp(0.632) = 1.881$$

Which gives the interval estimate of the *RR*: 1.393 (1.031-1.881)

The odds ratio is $$OR = \frac{166/1{,}176}{50/513} = 1.448$$

The standard error of the log(*OR*) is

$$SE(\log OR) = \sqrt{\frac{1}{166} + \frac{1}{1176} + \frac{1}{50} + \frac{1}{513}} = 0.1698$$

The lower and upper limits of the 95% CI of log(*OR*) are

$$L_{\log OR} = \log 1.448 - 1.96 \times 0.1698 = 0.038 \quad L_{OR} = \exp(0.038) = 1.038$$

$$U_{\log OR} = \log 1.448 + 1.96 \times 0.1698 = 0.703 \quad U_{OR} = \exp(0.703) = 2.020$$

Which gives the interval estimate of the *OR*: 1.448 (1.038-2.020)

***

3.2 Absolute measures of exposure effect

Relative measures alone may not provide a comprehensive information about the association between exposure and disease. As opposed to the *strength* of association estimated by relative measures, absolute measures show us the *impact* of the association between the exposure and the outcome of interest in public health terms: if the disease is more common, even exposures with lower relative effect may have more substantial absolute importance. The following absolute measures are often used for either risk (e.g., smoking) or protective (e.g., vaccination) factors (dos Santos Silva, 1999).

3.2.1 Risk factors

Risk difference (also called *excess risk* or *attributable risk*)

= risk in the exposed group – risk in the unexposed group

$$= R_1 - R_0 = \frac{a}{a+c} - \frac{b}{b+d}$$

Excess fraction (also called *excess risk percentage* or *attributable risk percentage*)

$$= RR - 1 = \frac{R_1}{R_0} - 1 = \frac{R_1 - R_0}{R_0}$$

3.2.2 Protective factors

To handle the preventive exposures, modification of the earlier measures was done by interchanging R_1 with R_0.

Risk reduction (also called *absolute risk reduction*)

= risk in the unexposed group – risk in the exposed group

$$= R_0 - R_1 = \frac{b}{b+d} - \frac{a}{a+c}$$

Prevented fraction (also called *relative risk reduction*; e.g., *vaccine efficacy*)

$$= 1 - RR = 1 - \frac{R_1}{R_0} = \frac{R_0 - R_1}{R_0}$$

4. Types of epidemiological studies

The studies in epidemiology can be either *observational* or *experimental* (Figure 2). In an observational study, the investigator purely measures the occurrence of exposures and/or outcomes. On the other hand, the principle of experimental studies is an active intervention to change a disease determinant, diagnostics, treatment, etc.

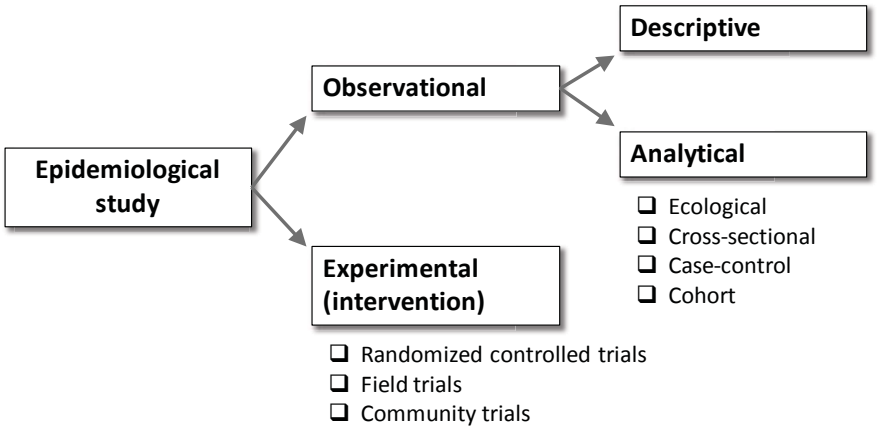


Figure 2. Types of epidemiological studies (classification from Bonita et al., 2006).

4.1 Observational studies

The simplest type of epidemiological study is a *descriptive study*, which means pure description of disease occurrence in the target population. This is often the first step in our epidemiological inquiry. Usually, we continue with an *analytical study*, which focuses on the quantification of relationships between the health status and a defined exposure.

Ecological study is usually useful as a hypothesis generation study. Rather than focusing on individual health outcomes and exposures, we are interested in aggregated figures for defined groups of people (e.g., in different regions or at different time periods). These studies are very simple and cheap (usually relying on data collected for different purposes); however, they tend to be difficult to interpret. Moreover, they may suffer from a serious bias (*ecological fallacy*), which stems from our inability to identify associations existing at the individual level.

Cross-sectional study measures both exposure and health status at the same time. Therefore, they are very useful to estimate the prevalence (of a disease or risk factor, they are also called prevalence studies); however, it is more difficult to draw causal inference from these studies, as we usually don't know whether the exposure preceded the effect.

Case-control studies (Figure 3) enable us to assess exposures and health status at different times (they are *longitudinal*). We start with the selection of *cases* (patients with the disease of interest among the entire target population). We also need to select *controls*; this step is critical when we perform case-control studies. The control group must sample the exposure prevalence in the target population. The exposure is usually evaluated retrospectively for

both cases and controls, e.g., by direct questioning, examining hospital records, or even by biochemical measurement. The association between the exposure and a disease is measured by calculating the odds ratio (*OR*). The case-control studies can be therefore used to estimate the relative risk of a disease, but we cannot estimate the prevalence of the disease, because it is defined by the epidemiologist conducting the study (Bonita et al., 2006).

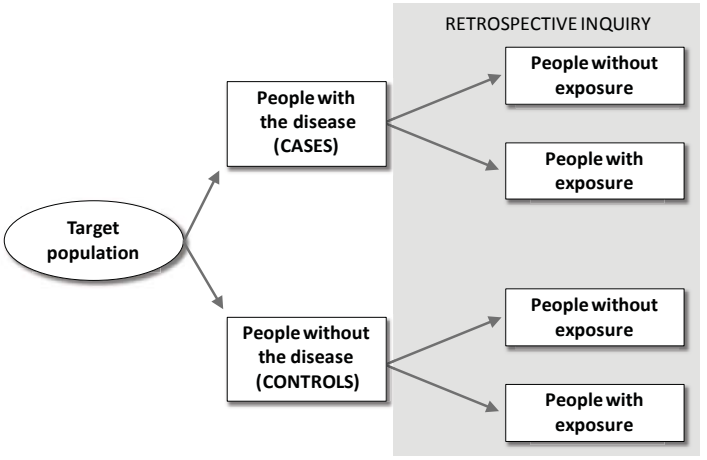


Figure 3. Scheme of a case-control study.

As opposed to the case-control studies, *cohort studies* (Figure 4) start with the selection of healthy individuals from the target population and with the ascertainment of their exposure status. The whole cohort is then followed up for a sufficient time interval to observe how many disease cases develop in both exposure groups. This gives them a longitudinal nature and indeed, cohort studies provide the best information among observation studies about the disease causation (Table 2). However, as it may take really a long time between the exposure and its effect, they may be both lengthy and costly.

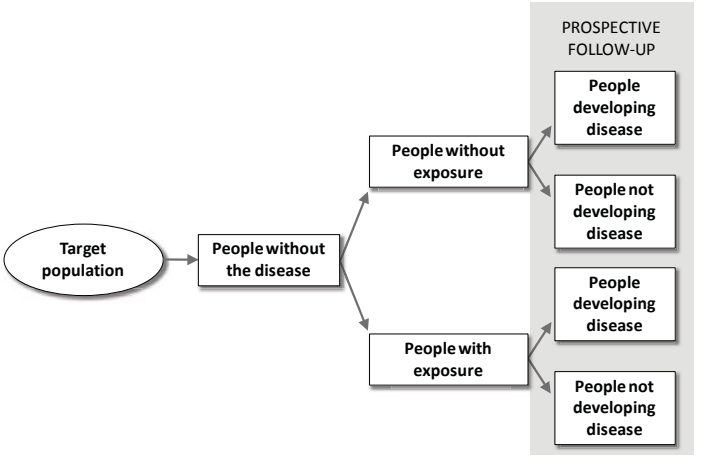


Figure 4. Scheme of a cohort study.

Table 2. Advantages and disadvantages of different observational study designs (Bonita et al., 2006).

Probability of	Ecological	Cross-sectional	Case-control	Cohort
selection bias	NA	medium	high	low
recall bias	NA	high	high	low
loss to follow-up	NA	NA	low	high
confounding	high	medium	medium	medium
time required	low	medium	medium	high
cost	low	medium	medium	high

NA: not applicable

4.2 Experimental studies

In experimental (intervention) studies, patients are assigned a particular exposure (treatment, prevention, etc.) by the researcher. The health outcomes are then compared in the experimental group (with the treatment) and the control group (without the treatment). The golden standard in the design of effectiveness studies are *randomised controlled trials*, where individuals are randomly allocated to the experimental or control group. *Field trials* usually take place among healthy people from the general population, who are perceived to be at risk of a disease, with the aim to prevent the disease. The study entity in *community trials* are not individuals, but entire communities who are allocated to treatment.

4.3 Sources of error in epidemiological studies

Random error is inevitable in studies based on population sampling. It stems from the individual biologic variation, sampling error and measurement error (Bonita et al., 2006). We can decrease the measurement error by stringent protocols and care taken during individual measurements. Individual variation always occurs in biological experiments; however, sampling error could be decreased by increasing the *sample size*. Sample size calculations can (and should) be done before conducting studies, so that we could be confident that the study is powerful enough to confirm the study hypothesis.

Systematic error (or bias) is a systematic difference of the study results from true values. The principal biases in epidemiology are selection and measurement biases. *Selection bias* means a systematic difference between the characteristics of the people selected for a study and the characteristics of those not selected. *Measurement bias* happens when measurements or classifications of individuals are inaccurate. A specific case of this bias – recall bias – occurs in the case-control studies when cases are more (or less) likely to recall some past exposure (Bonita et al., 2006). The third threat to validity of the study is called *confounding* and will be described in the next chapter.

5. Association vs. causation

5.1 Association and confounding

Ideally, we would like to know what would be the risk of a disease in a particular population under different conditions given by the presence or absence of an exposure – an exposure *effect* (Rothman et al., 2008). Although indicators outlined in section 2 are often called effect measures, in fact these are merely designed to capture an *association* between two variables. Consider an example where we estimate a ratio of lung cancer incidence in males and

females in the Czech Republic. Clearly, this is not the effect of changing sex in a particular population, merely a measure of association between the sex and incidence.

It is tempting to substitute association measures for effect measures and also to give them directly causal explanations. However, this approach may be incorrect. Let us think of an example where we try to find out the cause of lung cancer (Katz, 2006). We will consider carrying matches as a potential risk factor of the disease. We will perform the study and, indeed, find substantial association between carrying matches and the lung cancer.

The real question in epidemiology is, however, whether the lung cancer risk in a particular population would change if the individuals would or would not carry matches (in other words, would matches ban prevent lung cancer?). In this particular example, it is quite clear that matches themselves are not cause of the lung cancer and our causal explanation would be completely wrong. What is the reason for this error?

Take a look at the Figure 5. Of course, the real cause of lung cancer is smoking. The problem is that smoking is associated with carrying matches, but is not caused by carrying matches. Smoking induces lung cancer. We will therefore see a positive association between carrying matches and lung cancer, but causal interpretation of this association would be inappropriate. In such cases we call the real cause associated with a putative risk factor *confounding variable* (*confounder*).

These are the three necessary characteristics of a confounder (Rothman et al., 2008):

1. A confounding factor must be a risk factor for the disease.
2. A confounding factor must be associated with the exposure under study in the source population (the population at risk from which the cases are derived).
3. A confounding factor must not be affected by the exposure of the disease. In particular, it cannot be an intermediate step in the causal path between the exposure and the disease.

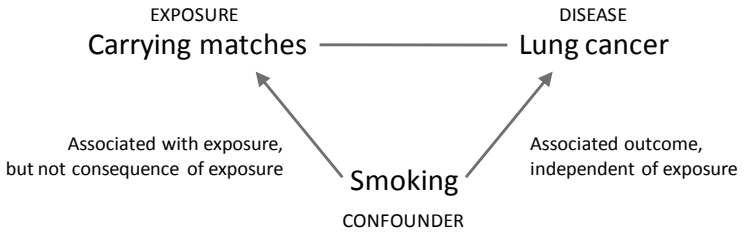


Figure 5. Causal diagram for lung cancer causation example.

5.2 Causation in epidemiology

Our aim in epidemiology is to prevent and control diseases and to promote health and, as emphasised above, we need to know the disease causes. *Cause* of a specific disease event may be defined as "an antecedent event, condition, or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are fixed" (Rothman and Greenland, 2005). Unfortunately, most identified causes are neither *necessary* nor *sufficient* to produce a disease; the interaction of many genetic and environmental factors is almost inevitably involved in any disease causation. However, removal of an identified risk factor may still lead to the prevention of a significant proportion of disease.

Before an association is assessed for the possibility that it is causal, other explanations, such as chance, bias and confounding, have to be excluded (Bonita et al., 2006). To have a framework to think about potential causes, different '*considerations for causation*' were proposed (e.g., Table 3) on the basis of aspects of association recommended by Hill (1965) to consider before deciding on potential causation.

Table 3. Considerations for causation (Bonita et al., 2006)

Temporal relation	Does the cause precede the effect? (essential)
Plausibility	Is the association consistent with other knowledge? (mechanism of action, evidence from experimental animals)
Consistency	Have similar results been shown in other studies?
Strength	What is the strength of the association between the cause and the effect? (relative risk)
Dose-response relationship	Is increased exposure to the possible cause associated with increased effect?
Reversibility	Does the removal of a possible cause lead to reduction of disease risk?
Study design	Is the evidence based on a strong study design?
Judging the evidence	How many lines of evidence lead to the conclusion?

It is necessary to acknowledge that the best evidence on causation could be provided by randomised controlled trials, as the randomisation is the best way to control confounding. However, such experiments are not always appropriate or even feasible (consider an example of a randomised controlled study assigning people to smoking or non-smoking). Therefore, observational studies have their important role in epidemiology; nevertheless, we need to bear in mind their potential risk of bias. Most notably, the ability of cross-sectional and ecological studies to confirm causation is rather weak. Of course, a study of any type must be well designed and performed before we can consider its results as a basis for causation.

6. Handling confounding in practice

6.1 Design phase

Randomisation is the essential element of randomised controlled trials. It means a random allocation of studied individuals to studied exposures (this is possible only in experimental studies). It enables us to create groups that should differ solely in the exposure under study; the distribution of other variables, notably confounders (both known and unknown), should be similar in both groups. We can therefore remove the association between exposure and confounders and thus prevent confounding.

Restriction is a very simple and effective way to limit confounding. If we can restrict the access to the study only to patients with a particular level of the selected confounder (e.g., particular race, age range, etc.), we prevent the confounder from varying and therefore exclude confounding. The main disadvantage of restriction is lowering the number of available subjects, possibly making the study unfeasible, when the resulting sample size is insufficient. Also, the generalization of the study results will be limited when using restricted study sample.

Matching refers to the selection of study participants in a reference series (i.e., unexposed subjects in a cohort study, controls in a case-control study) so that the selected subjects match the index series (exposed/cases) individuals with respect to one or several confounding factors values. Therefore, using matching, we are making the distribution of confounding factors similar in both study groups (Rothman et al., 2008).

6.2 Analysis phase

The simplest analysis for the assessment of confounding is *stratification*, i.e., creating separate estimates of effect for each group (*stratum)* of individuals according to values of the particular confounder.

Standardisation is taking a weighted average of stratum-specific outcomes (e.g., incidence rates) according to a defined standard – set of weights. The formula for a standardised rate is (Rothman et al., 2008):

$$I_w = \frac{\sum_i w_i I_i}{\sum_i w_i}$$

where w_i is the weight for stratum i and I_i is the rate in stratum i; w_i is usually the amount of person/time observed in stratum i of a standard population. Standardisation is often used to compare incidence rates between countries (using European or world standard population weights) with different age structures of population.

Supposing that stratum-specific estimates don't substantially differ from each other (such situation would suggest interaction – effect modification), estimates of exposure effect can be pooled using standard epidemiological techniques. *Pooled estimates* are weighted averages of stratum-specific estimates. As opposed to standardisation, pooling weights are applied to effect measures instead of occurrence measures and assume homogeneity of these effect measures between strata. Moreover, weights in pooling are internal and reflect the amount of information in each stratum (e.g., weights inversely proportional to stratum-specific estimates variance). As an example, consider computation of Mantel-Haenszel pooled estimate of *OR* (Rosner, 2006):

$$OR_{MH} = \frac{\sum_i \frac{a_i \cdot d_i}{n_i}}{\sum_i \frac{b_i \cdot c_i}{n_i}}$$

where a_i-d_i reflects number of patients in a 2×2 contingency table (like Table 1) created for the stratum i. For other estimators of pooled effect measures and associated variance see Rothman et al. (2008), for example.

The basic tabular methods are essential and often sufficient in epidemiologic data analysis; however, they fail if we need to examine many variables simultaneously. Under that condition, the usual method of choice is regression modelling. This topic will be addressed in detail in the following lesson.

***

Example 2. Hypothetical study examining the relationship between lung-cancer incidence and heavy drinking (Rosner, 2006):

COMPLETE STUDY GROUP		**Drinking status**		
		Heavy drinker	Nondrinker	Total
Lung cancer	Yes	33	27	60
	No	1,667	2,273	3,940
	Total	1,700	2,300	4,000

Odds ratio: 1.667 (95% CI 0.998-2.782).

The crude analysis suggests that drinking is a risk factor of lung cancer. Let us now investigate odds ratio separately for smokers and nonsmokers:

SMOKERS		**Drinking status**		
		Heavy drinker	Nondrinker	Total
Lung cancer	Yes	24	6	30
	No	776	194	970
	Total	800	200	1,000

Odds ratio: 1.000 (95% CI 0.403-2.480).

NONSMOKERS		**Drinking status**		
		Heavy drinker	Nondrinker	Total
Lung cancer	Yes	9	21	30
	No	891	2,079	2,970
	Total	900	2,100	3,000

Odds ratio: 1.000 (95% CI 0.456-2.192).

We can now see that the relationship between lung cancer and drinking was completely confounded by smoking (causally related to lung cancer and associated with drinking). Thus, the stratification disclosed that, in fact, drinking is not associated with lung cancer in neither smokers nor non-smokers. As we see that odds ratio is similar in both groups, we may proceed to pooling of results for different strata using Mantel-Haenszel estimator:

$$OR_{MH} = \frac{\frac{24 \cdot 194}{1000} + \frac{9 \cdot 2{,}079}{3000}}{\frac{776 \cdot 6}{1000} + \frac{891 \cdot 21}{3000}} = \frac{10.893}{10.893} = 1$$

***

7. References

dos Santos Silva I. Cancer epidemiology: principles and methods. Lyon: International Agency for Research on Cancer. 442 p. ISBN 92-832-0405-0

Bonita R, Beaglehole R, Kjellström T. Basic epidemiology. 2nd ed. Geneva: WHO, 2006. 213 p. ISBN 978-92-4-154707-9.

Hill AB. 1965. The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine 58:295-300.

Katz M. 2006. Multivariable Analysis. A practical Guide for Clinicians. 2nd ed. Cambridge: Cambridge University Press. 203 p. ISBN 978-0-521-54985-1.

Rothman KJ, Greenland S. 2005. Causation and causal inference in epidemiology. American Journal of Public Health 95: S144-50.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008. 758 p. ISBN 978-0-7817-5564-1.

Rosner B. Fundamentals of Biostatistics, 6th ed. Belmont: Thomson Higher Education, 2006. 868 p. ISBN 978-0-495-06441-1.

Saracci R. 2010. Introducing the history of epidemiology. In: Olsen J, Saracci R, Trichopoulos D (eds.). Teaching Epidemiology - A guide for teachers in epidemiology, public health and clinical medicine 3rd ed. Oxford: Oxford University Press, 2010. 576 p. ISBN 978-0-19-923947-4.

Woodward M. Epidemiology: study design and data analysis. Boca Raton: Chapman & Hall/CRC, 1999. 699 p. ISBN 1-58488-009-0.

Regression modelling in biomedical studies

Hynek Pikhart

Department of Epidemiology & Public Health, University College London;
e-mail: h.pikhart@ucl.ac.uk

Abstract

This contribution should help students to be able to:

- Identify variables which might be included in the statistical analysis using regression modelling
- Formulate a modelling strategy to build proper regression model
- Use a logistic model to compare the log odds of disease in 2 groups and to estimate a crude odds ratio for a binary outcome
- Perform statistical tests of the null hypothesis (= there is no association between the exposure and outcome)
- Use a logistic model with one exposure – either continuous or categorical with 2 or more levels
- Use a logistic model to examine the association between the exposure and outcome adjusting for confounders, assuming no effect modification, and explain the implications of such assumption
- Use likelihood ratio test in multiple regression models
- Use a logistic regression model that includes interaction parameters and interpret the parameters representing interaction in regression models

Key words

Epidemiology, regression modelling, multiple logistic regression, confounding, interaction

1. Strategies of the analysis

Introduction

In this section you should think about various issues covered in different parts of statistical courses and about practicalities of the analysis related to the range of methods and techniques which you have discussed previously. We consider what steps in the analysis you need to take, what techniques you should use at the beginning of your data analysis, and how to design and formulate modelling strategy (including decisions on possible confounders and effect modifiers)

Before you start the analysis

You should have clearly defined outcome and the main exposure (or exposures) in your data (this should be clear from your hypothesis/hypotheses). In such situation you know which variables are the main variables of your interest.

First step

You need to start with simple descriptive analysis – you should get to know your data, get the feeling for your data. You should firstly see what data are available in your dataset. You might then examine frequency distribution of the categorical data and you can try to graphically display your main variables. You can also look at the summary statistics of your continuous variables. Examining frequency distributions and graphs you should be able to identify possible errors in the data, find outliers in your data and see whether you have any missing data. This should also help you to decide how to categorise and/or regroup some variables.

Second step

Simple univariate analysis should be followed by bivariate analysis. Simple cross tabulations of two variables will give you the feeling for the crude associations in the dataset and will allow you to see how many events (cases of disease, deaths) you have in each category of exposure. The cross-tabulations may also help to find further potential errors in the dataset.

Third step

Use Mantel-Haenszel method for identification of possible confounders and effect modifiers. M-H method has possibly less power than regression technique however it gives you clearer results – it gives you stratum specific estimates in addition to overall pooled result, and it also gives you the test for effect modification.

Final step

Only as a final step, you should use regression modelling. By now you should have identified potential confounders and potential effect modifiers and you need to evaluate their effects in more complex models (than those available in M-H statistical technique). You need to consider whether the variable has any effect on the outcome, whether the variable has any effect on the association between main exposure(s) and the outcome. Before considering variable to be confounder you need to test whether such variable does not act as an effect modifier in the association between main exposure(s) and the outcome. In large datasets, with large number of possible confounders, you need to consider which variables should be included in regression model, you need to assess the associations between potential confounders and effect modifiers, and you need to assess potential dose-response effects of variables.

2. Logistic regression

Introduction

Logistic regression is more general method for analysis of binary outcomes than chi-square method since it allows the inclusion of continuous explanatory variables, inclusion of more than one exposure and the assessment of interaction (effect modification) between variables.

Revision – odds and odds ratio

Odds of disease and odds ratio can be defined in following way:

For a defined population and time period, it is the number of cases divided by the number of people who did not became a case

$$\text{Odds} = \frac{\text{Cases}}{\text{Non cases}} = \frac{\text{number with the disease}}{\text{number without the disease}}$$

Example of odds:

	CVD
No	18,954
Yes	2,676
Total	*21,630*

The odds of CVD is calculated as: Odds = 2,676/18,954 = 0.14

The examples shown in this section were calculated in statistical package Stata but most statistical packages can calculate similar outcomes. We are not interested in the syntax of commands – we need to focus on interpretation of results.

cases	controls	odds	[95% Conf.	Interval]
2676	18954	0.14118	0.13558	0.14702

Same odds as the one above calculated by hand

Now, let's calculate odds ratio

$$\text{Odds ratio (OR)} = \frac{\text{Odds in exposed group}}{\text{Odds in unexposed group}}$$

Example:

	TV watching		
Obesity	**<3 hours a day**	**>= 3 hours a day**	**Total**
Non obese	1,270	527	1,797
Obese	409	219	628
Total	1,679	746	2,425

OR= $Odds_{exp}$/ $Odds_{unexp}$

- $Odds_{unexp}$ =409/1270=0.32 (odds of obesity among those watching TV <3hours a day)
- $Odds_{exp}$ =219/527=0.42 (odds of obesity among those watching TV >=3hours a day)
- OR = 0.42/0.32
- OR =1.29

And in STATA:

```
mhodds obesity tv
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf.	Interval]
1.290369	6.72	0.0095	1.063563	1.565541

Comparing: >=3hours a day of TV vs <3 hours a day

We get the same odds ratio as in our calculation by hand

However we get extra information (which we could calculate by hand as well):

We get confidence interval – in this case 95% confidence interval – 1.06-1.57

And we also get p value related to the appropriate null hypothesis – p=0.0095

Exercise: Can you interpret these values?

Let's return to odds ratios. We can express formula in other way:

```
Odds in exposed group = Odds in unexposed group  x  Odds ratio(of exposure)
```

We can now logarithm the formula and get

```
Log(Odds in exposed group) = log(Odds in unexposed group) + log(Odds ratio)
```

Logistic regression model

Modelling log odds is referred to as logistic regression, and the models are named as logistic models.

Why do we use log odds when fitting statistical model: The reason for modeling log odds rather than risk or odds is that log odds can take any value (negative or positive) while risk lies between 0 and 1 and odds lies between 0 and infinity. When using statistical model it is easier to model a quantity which is unconstrained (which avoids the possibility to predict impossible values).

If we come back to our formula:

Log ($odds_{exp}$)=Log($odds_{unexp}$)+log(OR)

We will call **Log($odds_{unexp}$)** as **baseline** (log odds of disease in the unexposed group) and **log(OR)** as the effect of **exposure** (our main interest).

In summary, in logistic regression we fit regression model (with intercept and slope) for the log odds of disease as the outcome measure.

The model is fitted using a mathematical technique called maximum likelihood which takes into account that the variation of proportion has a binomial distribution.

A logistic regression with a binary exposure variable

Example

We want to see whether sex is risk factor for the all-cause mortality in population-based study:

	mortality		
	0	1	Total
women 0	119	31	150
	79.33%	20.67%	100.0%
men 1	239	131	370
	64.59%	35.41%	100.0%
Total	358	162	520
	68.85%	31.15%	100.0%

Our outcome is all-cause mortality. Our exposure is gender. Let's calculate odds of the outcome in both genders.

Odds (women) = 31 / 119=0.26

Odds (men) = 131 / 239 = 0.548

OR (men vs women) = 0.548 / 0.26 = 2.11

Now, we can calculate logistic regression in our statistical package. In Stata, output would look like this:

```
logit mort i.sex
```

```
Logistic regression                                     Number of obs   =        520
                                                        LR chi2(1)      =      11.35
                                                        Prob > chi2     =     0.0008
Log likelihood = -316.89647                             Pseudo R2       =     0.0176

------------------------------------------------------------------------------
        mort |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   .7438701   .2290832     3.25   0.001     .2948753    1.192865
       _cons |  -1.345136   .2016468    -6.67   0.000    -1.740357   -.9499159
------------------------------------------------------------------------------
```

The constant refers to the log odds in the baseline group. The coefficient gives the Maximum likelihood estimate of the parameter.

ln odds= -1.345136+0.7438701 x sex (0"women" 1"men")

ln odds (unexposed=women) = -1.345+0.744 x 0= -1.345

odds (unexposed) = exp(-1.345)=0.26

ln odds (exposed=men) = -1.345+0.744 x 1= -0.601

odds (exposed) = exp(-0.601)=0.548

OR = odds(exposed)/odds(unexposed) = 0.548/0.26 = 2.10

If we take the estimate from the STATA model, 0.74387, we will see that

exp (0.74387) = 2.10

STATA can provide the Odds Ratios (OR) which are more familiar and easy to interpret.

```
logistic mort i.sex
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(1)      =      11.35
                                                  Prob > chi2     =     0.0008
Log likelihood = -316.89647                       Pseudo R2       =     0.0176
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   2.104063   .4820054     3.25   0.001     1.342959    3.296512
------------------------------------------------------------------------------
```

We can see that we got the same OR as in previous calculation.

Similarly, we can calculate 95% confidence interval from logit model:

95%CI for OR = exp(0.2948) to exp (1.1928) = 1.34 to 3.30 which is identical to the 95% CI from logistic model.

Finally, we can use Z statistic (that can be compared with a Normal distribution) for significance testing of the strength of the association:

We use the Wald test to test the null hypothesis that the true parameter value is 0 (i.e., there is no association)

z statistic is calculated as

Z = coefficient/SE

Z = ln (OR) / SE (lnOR) Here we must use coefficient and SE from original, logit model!

We compare z with a Normal distribution

For our example

Z = 0.744/0.229 = 3.25

p=0.001 we reject the null hypothesis of no association

Testing for association using the Likelihood ratio test

For each logit regression model you can calculate "log likelihood" statistics

The Likelihood Ratio Test (LRT) –

- LRS (likelihood ratio statistics) = 2 (L1-L0), where L1 is maximised log likelihood of model with variable you want to test and L0 is maximised log likelihood of model without the variable
- LRS is distributed as chi-square distr. on 1 df (if we test effect of 1 variable, later we will try to test composite effect of more variables)

STATA:

```
logistic mort i.sex        * more complicated model
i.sex             _Isex_0-1          (naturally coded; _Isex_0 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(1)      =      11.35
                                                  Prob > chi2     =     0.0008
Log likelihood = -316.89647                       Pseudo R2       =     0.0176
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   2.104063   .4820054     3.25   0.001     1.342959    3.296512
------------------------------------------------------------------------------

est store a                        * store estimates

logistic mort              * less complicated model

Logistic regression                               Number of obs   =        520
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -322.56957                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------

est store b                        * store estimates

lrtest b a                         * compare estimates of models a and b

Likelihood-ratio test                                 LR chi2(1)  =     11.35
(Assumption: b nested in a)                           Prob > chi2 =    0.0008
```

When we compare model including variable sex and model without this variable, Likelihood ratio test again shows importance of sex on mortality in this dataset.

Important points:

- LRT can be used even in more complicated situations (we will see later)
- We can only use LRT test if both compared models have same number of individuals used in regression analysis (you must check that there are no missing values in variable(s) tested!)
- Two compared models must be nested (exposures used in less complicated model are subset of exposures used in more complicated model)

Logistic regression for the comparison of more than 2 groups

We have categorical exposure that has more than two categories. Let's come back to our example and use variable age. We have 3 age groups (50 years and younger, 51-65, older than 65) and we want to see the effect of age (grouped to 3 categories) on all-cause mortality.

Firstly, let's tabulate two variables:

agegp	mortality 0	1	Total
1	142	16	158
	89.87%	10.13%	100.00%
2	116	56	172
	67.44%	32.56%	100.00%
3	100	90	190
	52.63%	47.37%	100.00%
Total	358	162	520
	68.85%	31.15%	100.00%

We can see that age seems to be associated with mortality (10% of dead individuals in the youngest age group, 33% in the middle group and 47% in the oldest group). Let's run logistic regression:

```
logistic mort i.agegp
i.agegp           _Iagegp_1-3         (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(2)      =      61.60
                                                  Prob > chi2     =     0.0000
Log likelihood = -291.76873                       Pseudo R2       =     0.0955
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |   4.284483   1.327637     4.70   0.000     2.334187    7.864319
   _Iagegp_3 |     7.9875   2.404932     6.90   0.000      4.42716    14.41108
------------------------------------------------------------------------------
```

Age 51-65 vs 50 and less **Age 65+ vs 50-**

_Iagegp_2 and _Iagegp_3 are **indicator variables** created by STATA for each non-baseline value of categorical variable for the purposes of analysis. Indicator variable take only values 0 and 1.

_Iagegp_2 is an indicator variable that equals 1 for agegp=2 and equals 0 otherwise

_Iagegp_3 is an indicator variable that equals 1 for agegp=3 and equals 0 otherwise

So, the 4.28 is the odds ratio comparing individuals in **age group 2** (51-65 years) vs those in **age group 1** (baseline; 50 and less). The remaining columns have the same meaning as previously, so we can see that 95% CI for OR is 2.33-7.86, and OR is statistically significantly different from 1.00 (no association).

Similarly, 7.99 is the odds ratio comparing individuals in **age group 3** (65+) vs those in **age group 1** (baseline)!

The estimated OR always compares appropriate category of the variable with the baseline!

So far, we have tested whether mortality in age group 2 differs from mortality in age group 1 and whether mortality in age group 3 differs from mortality in age group 1. No we are interested in **composite effect of age.** In other words, we want to know whether age is statistically associated with mortality. We need to use likelihood ratio test.

```
logistic mort i.agegp          * more complicated model
i.agegp           _Iagegp_1-3        (naturally coded; _Iagegp_1 omitted)
Logistic regression                          Number of obs   =        520
                                             LR chi2(2)      =      61.60
                                             Prob > chi2     =     0.0000
Log likelihood = -291.76873                  Pseudo R2       =     0.0955
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |   4.284483   1.327637     4.70   0.000     2.334187    7.864319
   _Iagegp_3 |     7.9875   2.404932     6.90   0.000      4.42716    14.41108
------------------------------------------------------------------------------

est store a

logistic mort              * less complicated model
Logistic regression                          Number of obs   =        520
                                             LR chi2(0)      =      -0.00
                                             Prob > chi2     =          .
Log likelihood = -322.56957                  Pseudo R2       =    -0.0000
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------

est store b
lrtest b a                     * likelihood ratio test
Likelihood-ratio test                        LR chi2(2)  =     61.60
(Assumption: b nested in a)                  Prob > chi2 =     0.0000
```

2xdiff between log likelihoods

We should repeat several basic points:

- 2 models must be nested

- Same number of subjects in both models

- degrees of freedom = 2

  = equal to difference in number of variables between 2 models

  (we had 2 dummy variables for agegroup, so 2 d.f.)

Logistic regression with quantitative measure of exposure

We can have continuous variable as exposure (systolic blood pressure, blood cholesterol, height). We want to estimate the effect of continuous exposure on binomial outcome.

We will use diastolic blood pressure (DBP) in our example:

Firstly, let's check whether we have DBP values for all individuals and whether we do not have any outliers (unusually small or large values):

```
sum dbp

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         dbp |       520    86.37308      8.8363         73        107
```

We can see that we have reasonable values for all 520 subjects. We can now run regression model:

```
logistic mort dbp

Logistic regression                                     Number of obs   =        520
                                                        LR chi2(1)      =      90.14
                                                        Prob > chi2     =     0.0000
Log likelihood = -277.50184                             Pseudo R2       =     0.1397

------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dbp |   1.111366   .0133546     8.79   0.000     1.085497    1.137851
------------------------------------------------------------------------------
```

The estimate of the exposure effect on mortality per **1 unit increase** in DBP

We can say that odds of mortality increases 1.11-times with 1 unit increase in diastolic blood pressure. OR=1.11 represents the effect of DBP per 1 mmHg increase in DBP. We can now use 95% CI and p-value in the same way as in previous examples.

If we want to estimate OR for 10 units increase in DBP the effect will be $(1.11)^{10} = 2.83$

3. Multiple logistic regression

Logistic regression allows using several confounding variables at the same time, allows inclusion of possible effect modifiers and allows using continuous variables as confounding factors.

Adjusting for confounding using multiple logistic regression

Let's return to our example (sex, age and diastolic blood pressure as possible risk factors for all-cause mortality)

We want to fit a logistic regression model including terms for both sex and age group at the same time.

We can use following STATA command

```
. xi: logistic mort i.agegp i.sex
```

We list both exposures in the command

```
i.agegp           _Iagegp_1-3         (naturally coded; _Iagegp_1 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |    6.28714   2.014746     5.74   0.000     3.354913    11.78216
   _Iagegp_3 |   11.12581   3.463072     7.74   0.000     6.044813    20.47768
     _Isex_1 |   3.512644   .8555701     5.16   0.000     2.179256    5.661874
------------------------------------------------------------------------------
```

How should we interpret such results?

The parameter estimate for sex (odds ratio 3.51) represents the odds ratio for the effect of sex (men vs women) **adjusted for any confounding effect of age group**. In simple way we can imagine that we create separate tables and calculate odds ratios of the effect of sex on

mortality for each age group, and we make pooled estimates ~ weighted average of stratum specific odds ratios.

The age parameters can be interpreted in similar way: ORs of 6.29 and 11.12 represent the odds ratios for the effect of age (51-65 vs 50- and 65+ vs 50-) on all-cause mortality adjusted for any confounding effect of sex.

We need to mention one important assumption – we assume that there is no interaction/effect modification between the effects of age group and sex (M-H methods provides us with reminders about the effect modification while logistic regression does not). In other words we assume that the effects of sex and age group on mortality are independent (or, in other words, we assume that the effect of sex on all-cause mortality is same in all categories of age and the effect of age on all-cause mortality is same in both genders).

Hypothesis testing in multiple logistic regression

We can test different hypotheses in multiple logistic regression.

a) the composite effect of age on mortality (when sex taken into account)

We want to test following null hypothesis: there is no association between age group and mortality after taking sex into account. We will use likelihood ratio test for testing this hypothesis. We will use similar set of commands as in last session.

```
. xi: logistic mort i.agegp i.sex          * more complicated model
i.agegp           _Iagegp_1-3         (naturally coded; _Iagegp_1 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(3)      =      90.89
                                                  Prob > chi2     =     0.0000
Log likelihood =  -277.1232                       Pseudo R2       =     0.1409
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |    6.28714   2.014746     5.74   0.000     3.354913    11.78216
   _Iagegp_3 |   11.12581   3.463072     7.74   0.000     6.044813    20.47768
     _Isex_1 |   3.512644   .8555701     5.16   0.000     2.179256    5.661874
------------------------------------------------------------------------------
. est store a

. xi: logistic mort i.sex                  * less complicated model
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(1)      =      11.35
                                                  Prob > chi2     =     0.0008
Log likelihood = -316.89647                       Pseudo R2       =     0.0176
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   2.104063   .4820054     3.25   0.001     1.342959    3.296512
------------------------------------------------------------------------------

. est store b
. lrtest b a
Likelihood-ratio test                             LR chi2(2)  =     79.55
(Assumption: b nested in a)                       Prob > chi2 =    0.0000
```

The result of the likelihood ratio test tells us that there is very strong evidence against the null hypothesis ($p<0.001$) – there is strong evidence that, taking sex into account, there is an association between age group and odds of death.

This LRT test tells us whether there is evidence that a variable is **a risk factor** – it is **not** a test for whether variable is a **confounder**!

b) the composite effect of age and sex on mortality

This time, we want to test following hypothesis: there is no composite effect of age group and sex on mortality. We will again use likelihood ratio test for testing this hypothesis but we will compare different models:

```
. xi: logistic mort i.agegp i.sex          * more complicated model
i.agegp           _Iagegp_1-3         (naturally coded; _Iagegp_1 omitted)
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(3)      =      90.89
                                                  Prob > chi2     =     0.0000
Log likelihood =  -277.1232                       Pseudo R2       =     0.1409
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |    6.28714   2.014746     5.74   0.000     3.354913    11.78216
   _Iagegp_3 |   11.12581   3.463072     7.74   0.000     6.044813    20.47768
     _Isex_1 |   3.512644   .8555701     5.16   0.000     2.179256    5.661874
------------------------------------------------------------------------------

. est store a

. xi: logistic mort                       * less complicated model
Logistic regression                               Number of obs   =        520
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -322.56957                       Pseudo R2       =    -0.0000
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------

. est store b
. lrtest b a
Likelihood-ratio test                             LR chi2(3)  =      90.89
(Assumption: b nested in a)                       Prob > chi2 =     0.0000
```

The result of the likelihood ratio test tells us that there is very strong evidence against the null hypothesis ($p<0.001$) – there is strong evidence that there is **composite** effect of sex and age on all-cause mortality.

This type of hypothesis testing is particularly useful when we have blocks of variables of similar type or origin (for example several SES measures or several health behaviours) and we want to test their composite effect on health outcome of interest.

Interaction in logistic regression

So far, we needed to make the assumption that the effect of the exposure is the same (or similar) across the strata (=for different categories of confounder). We need to test such assumption in regression model (you may remember test for heterogeneity of odds ratios in Mantel-Haenszel analysis).

Let's return to our example. For simplicity, let's combine people older than 60 years into one group = we will have only 2 age groups. We want to test whether the effect of age group on

mortality is same among men and women (we want to test whether stratum-specific ORs are homogenous or not.

As always, we will construct 2-way tables first:

MEN:

agegp	mortality 0	1	Total
1	128	38	166
	77.11	22.89	100.00
2	111	93	204
	54.41	45.59	100.00
Total	239	131	370
	64.59	35.41	100.00

WOMEN:

agegp	mortality 0	1	Total
1	66	2	68
	97.06	2.94	100.00
2	53	29	82
	64.63	35.37	100.00
Total	119	31	150
	79.33	20.67	100.00

We can calculate stratum-specific odds ratios

```
. xi:logistic mort i.agegp if sex==1
```

We specify that regression will be conducted only among men

```
i.agegp           _Iagegp_1-2         (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        370
                                                  LR chi2(1)      =      21.12
                                                  Prob > chi2     =     0.0000
Log likelihood =  -229.9087                       Pseudo R2       =     0.0439
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |   2.822191   .6551494     4.47   0.000     1.790551    4.448218
------------------------------------------------------------------------------
```

```
. xi:logistic mort i.agegp if sex==0
```

Now, we specify that we will use only women

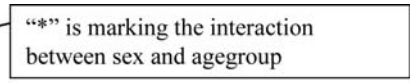
```
i.agegp           _Iagegp_1-2         (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        150
                                                  LR chi2(1)      =      28.26
                                                  Prob > chi2     =     0.0000
Log likelihood = -62.296944                       Pseudo R2       =     0.1849
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iagegp_2 |    18.0566   13.61451     3.84   0.000      4.11944     79.1469
------------------------------------------------------------------------------
```

MEN: OR= 2.82 WOMEN: OR= 18.06

These two ORs do not seem to be similar but we need to test this difference formally – it is possible for example that this difference is seen just because there are relatively few younger women in the sample who have already died (and we can see that 95% CI for the OR in women is extremely wide).

We need formal test of null hypothesis: stratum specific ORs are homogenous (there is no difference between stratum specific odds ratios)

Firstly we run the more complicated model = **model assuming interaction** between age and sex = model assuming that the effect of age on mortality depends on sex (and also assuming that the effect of sex on mortality depends on age)




```
. xi:logistic mort i.sex*i.agegp
i.sex             _Isex_0-1         (naturally coded; _Isex_0 omitted)
i.agegp           _Iagegp_1-2       (naturally coded; _Iagegp_1 omitted)
i.sex*i.agegp     _IsexXage_#_#     (coded as above)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(3)      =      60.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -292.20564                       Pseudo R2       =     0.0941
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   9.796875   7.260721     3.08   0.002     2.292133    41.87312
   _Iagegp_2 |    18.0566   13.61437     3.84   0.000       4.1195    79.14576
_IsexXage_~2 |   .1562969   .1233043    -2.35   0.019     .0332988    .7336222
------------------------------------------------------------------------------

. est store a
```

Then we ran simpler model = model assuming no interaction between age and sex

```
. xi:logistic mort i.sex i.agegp
i.sex             _Isex_0-1         (naturally coded; _Isex_0 omitted)
i.agegp           _Iagegp_1-2       (naturally coded; _Iagegp_1 omitted)

Logistic regression                               Number of obs   =        520
                                                  LR chi2(2)      =      52.84
                                                  Prob > chi2     =     0.0000
Log likelihood =  -296.1497                       Pseudo R2       =     0.0819
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   2.213595   .5250567     3.35   0.001      1.39058    3.523709
   _Iagegp_2 |    3.70086    .789548     6.13   0.000      2.43616    5.622112
------------------------------------------------------------------------------

. est store b
```

Finally we use likelihood ratio test to compare these two models

```
. lrtest b a

Likelihood-ratio test                             LR chi2(1)  =       7.89
(Assumption: b nested in a)                       Prob > chi2 =     0.0050
```

The result of likelihood ratio test tells us that there is evidence against null hypothesis (p=0.005) and we should not use model assuming independent effect of age and sex on mortality = we should report stratum specific odds ratios of the effect of age and sex on mortality:

```
. xi:logistic mort  i.sex*i.agegp
------------------------------------------------------------------------------
        mort | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_1 |   9.796875   7.260721     3.08   0.002     2.292133    41.87312
   _Iagegp_2 |    18.0566   13.61437     3.84   0.000       4.1195    79.14576
_IsexXage_~2 |   .1562969   .1233043    -2.35   0.019     .0332988    .7336222
------------------------------------------------------------------------------
```

Interpretation:

- Among younger people (60 years or less): the odds ratio for the effect of gender (men vs women) is 9.80 (_Isex_1)
- Among women: the odds ratio for the effect of age (older vs younger) is 18.06 (_Iagegp_2)
- Among older people (above 60): the odds ratio for the effect of gender (men vs women) is 9.80 multiplied by 0.156 (IsexXagegp_2) = 1.53
- Among men: the odds ratio for the effect of age is 18.06*0.156 = 2.82

Interpreting interaction terms

If there is important interaction in the model, it does not make sense to report the effect of the exposure on the outcome adjusted for confounder – the proportional odds assumption is not correct. We must report stratum-specific exposure effects (in both directions of interaction).

We will continue with confounding and interaction in multiple logistic regression case study session.

Public sources of cancer epidemiology

Jan Mužík, Tomáš Pavlík, Ladislav Dušek

Institute of Biostatistics and Analyses, Masaryk University, Brno; e-mail: muzik@iba.muni.cz

Abstract

Cancer epidemiology can be regarded as one of the most important and most frequently analyzed topics in the field of human risk assessment. Demand for cancer epidemiology data cannot be easily fulfilled by blind outputs bearing only primary population-based data. Therefore, development of the professional information sources providing the data in user friendly and accessible form is required. Information sources are nowadays available in the form of analytical tools, which enable access to different types of data sources on national and international level. In the Czech Republic it is System for Visualization of Oncology Data (SVOD), on the international level it is CANCERMondial by International Agency for Research on Cancer (IARC), which joins access to different international sources dealing with cancer epidemiology. Thanks to these tools is the information on cancer epidemiology widely and easily accessible.

Key words

Cancer, epidemiology, information tools.

1. Introduction

Cancer epidemiology can be regarded as one of the most important and most frequently analyzed topics in the field of human risk assessment. It is not only due to remarkable public concern about the growing population risk; cancer incidence and mortality are evident and clearly attainable endpoints for risk and health care assessment studies. We can enter this field from the viewpoint of risk factors as agents initiating carcinogenesis, but epidemiological parameters can also retrospectively indicate hazardous population impact on a large scale. However, the indication based on epidemiological data of course requires sufficient data sources. It means having representative long-term profiles of incidence and mortality as well as very good awareness of most important risk factors.

We need easily available large data sets, which themselves are, however, very expensive and typically not directly available. That is why we must aggregate at least cancer and demographic data in order to attain relevant age-adjusted profiles of epidemiologic parameters. It is also the main reason for growing interest in accessibility of population-based data, recently expressed by many professional groups (health care managers, environmental experts, risk assessors). According to our experience, however, their demand for data cannot be easily fulfilled by blind outputs bearing only primary population-based data. Therefore, development of professional web portals that offer automatically generated and verified epidemiological analyses on cancer incidence and mortality is needed.

2. Sources of information on cancer epidemiology

In the background of the information sources on cancer epidemiology there are three crucial types of population-based data:

- cancer incidence data – collected in cancer registries
- cancer mortality data – collected in specific databases of deceased persons or in cancer registries
- data on population structure – collected by governmental statistical offices

All these data sources should be representative for the population of interest; moreover, cancer registries should fulfil international recommendations and criteria on data structure and data completeness. On the basis on such reliable population-based data sets, comprehensive information sources on cancer epidemiology on regional, national or international level can be developed.

2.1. Sources of cancer epidemiology data in the Czech Republic

The crucial data source for cancer epidemiology in the Czech Republic is Czech National Cancer Registry (CNCR), which is managed and guaranteed by the Institute of Health Information and Statistics of the Czech Republic (IHIS CR) at the Czech Ministry of Health. Standardized collection of cancer data in CNCR started in 1977 and provides representative long-term trends for most of the cancer diagnostic groups. Nowadays, the database consists of more than 1.8 million cases stratified according to main risk factors and diagnostic descriptors including TNM classification of tumours. Basic outputs of CNCR are available in annual reports of IHIS CR (IHIS CR, 2013). To make these unique data source accessible for specialists and also for public in user friendly form, unique automated system of on-line analyses was developed. This system is located at the web portal SVOD (System for Visualization of Oncology Data), which is available at http://www.svod.cz (Dusek et al., 2005).

2.1.1. SVOD - System for Visualization of Oncology Data

The portal SVOD (System for Visualization of Oncology Data) is aimed to provide user-controlled analyses over available data sources (e.g. cancer epidemiology, demographic data). All analytic functions are accompanied with proper visualisation – graphical and table protocols that can be further exported and used. The portal functions are targeted primarily for health care managers and risk assessors working in the field of human and ecological risk assessment, but all outputs are designed to be widely accessible to general public. Analytical tools available at http://www.svod.cz can be summarized as follows:

- ***Incidence and mortality***: time trends of incidence, mortality and mortality/incidence ratio. Available parameters are absolute numbers of incident cases, crude rate (number of cases per 100,000 persons in population) and age standardized ratio (ASR - European or World age standard)
- ***Time trends***: changes of incidence and mortality in time. Available parameters are growth indices related to selected year and between-years changes. Both parameters could be viewed as absolute numbers or as relative percents.
- ***Regional overview***: comparison of incidence and mortality in regions of the Czech Republic. Available parameters are crude rate and age standardized ratio (ASR - European or World standard).

- ***Age-adjusted analyses***: age structure of population of patients with selected diagnosis.
- ***Clinical stages***: time trends in proportion of patients diagnosed in a specific clinical stage. Available parameters are absolute numbers, percents and crude rate of patients in specific clinical stage(s).
- ***International data***: comparison of incidence and mortality in the Czech Republic with other countries. All these analyses are based on data obtained from IARC database GLOBOCAN 2008.
- ***Comparative standards***: time trend of incidence or mortality in selected region in comparison with situation in the Czech Republic.
- ***Typology of patients***: comprehensive overview of group of patient with specified diagnosis.

Each output can be modified by selection of specific settings of the analysis (units, viewed parameters, type of graph etc.) or by selection of a specific group of patients according to sex, age group, region, time period, clinical stage, TNM classification and other parameters related to the status of the patient. Examples of automated analytic tools and outputs can be seen on figure 1.

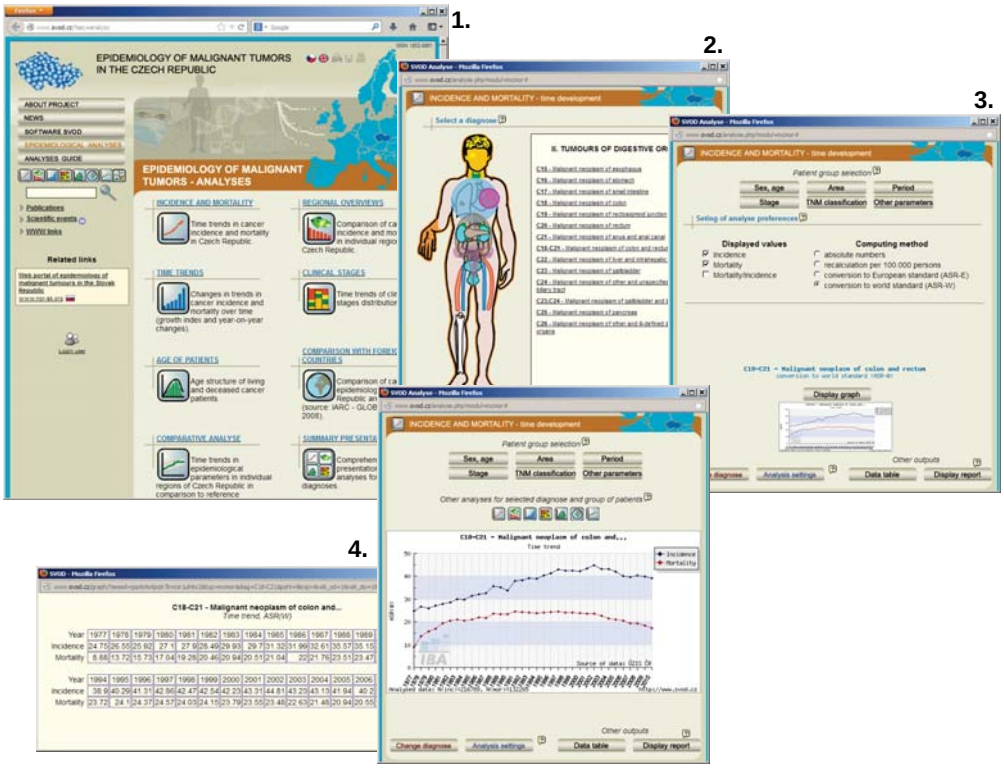


Figure 1. Selection of specific output in four steps: 1. selection of specific analytic tool; 2. selection of diagnosis of interest; 3. analysis setting and selection of target group of patients; 4. graphic and tabular outputs and reports.

2.2. Sources of international cancer epidemiology data

There are many internet sites, which provide information on cancer epidemiology on regional or national level of specific countries and regions. Nevertheless, most of the users mainly need a comprehensive information source comparing cancer data from different countries and regions at one site. Such types of sources are available and are grouped on the site of International Agency for Research on Cancer (IARC) called "CANCERMondial" (http://www-dep.iarc.fr). CANCERMondial provides access to specific individual international information sources of cancer incidence, mortality, prevalence and survival of specific cancers. These sources are specific in the type of information provided and the most important of them will be described in following chapters.

2.2.1. Cancer Incidence in Five Continents

The Cancer Incidence in Five Continents (CI5) series of monographs, published every five years, has become the reference source of data on the international incidence of cancer. The on-line CI5 databases provide access to detailed information on the incidence of cancer recorded by cancer registries (regional or national) worldwide in two formats:

1. **I5 I-IX** which presents the data published in the nine volumes of CI5
2. **CI5plus** which contains annual incidence for selected cancer registries published in CI5 for the longest possible period

The first format (I5 I-IX) provides tabular outputs of data, which were published in individual volumes of printed CI5. The second format (CI5plus) is more sophisticated and provides tabular outputs by populations, cancers or years and graphic outputs in the form of age-specific curves, time trends, time trends by age and trends by birth cohort. Moreover, the data sources, from which all outputs of CI5 are calculated, are available for download.

2.2.2. WHO Cancer Mortality Database

This database, created and maintained by the Section of Cancer Information at IARC, contains selected cancer mortality statistics by country, extracted from the World Health Organisation (WHO) database. The original data have been converted and/or recoded to a common system before presentation. However, due to changes in the ICD overtime, limited number of cancer sites is available and some parts are incomplete, particularly from earlier time periods.

Outputs are available in tabular form (by cancers, populations or years) or in graphic form: line charts or age-specific curves, cumulative risk by age, time trends, time trends by age and time trends by birth cohort; bar chart comparing different populations; pie chart of cancer mortality proportions in selected population; population pyramid – age and sex structure of selected population. Analytic tools also provide short and long term predictions of cancer mortality and identification of single break points in mortality trends (so called significant change in trends).

2.2.3. GLOBOCAN 2008

GLOBOCAN 2008 represents more comprehensive information source, which is based on raw data from cancer and mortality databases. GLOBOCAN provides access to the most recent **estimates** of the incidence, mortality, prevalence and disability-adjusted life years (DALYs) for major type of cancers at national level for 184 countries of the world. The recent GLOBOCAN estimates are presented for 2008, separately for each sex and, for incidence and mortality data, for ten age groups. One-, three- and five-year prevalence data

are available for the adult population only (age 15 and more). These estimates are based on the most recent data available at IARC and on information publicly available on the Internet, but more recent figures may be available directly from local sources. Due to continuous improving in quality and extent of data sources used for GLOBOCAN calculations, the outputs are relatively frequently updated a corrected.

The GLOBOCAN outputs are available in a form of complete factsheets, which describe the overview of the selected type of cancer or cancer epidemiology in the selected country, or as graphs, maps and tables focused on specific topic (e.g. incidence of specific cancer in men in different countries). Following outputs for incidence/mortality are available:

- Tables: age-specific rates or numbers; standardised rates by populations or cancers
- Graphs: age-specific incidence/mortality curves; multi-bar chart by populations or by cancers; dual multi-bar chart by cancers /populations, by cancers/sexes or by populations/sexes; cancer maps; pie chart by populations or by cancers; population pyramid by age/sex
- Advanced option: predictions of incidence (nowadays for 2008 and 2010) in graphs and tables

Prevalence estimates (1-, 3- and 5-year) and disability-adjusted life years are available as tables: proportions by populations or by cancers and graphs: dual multi-bar chart (showing incidence and prevalence) by cancers /populations, by cancers/sexes or by populations/sexes; cancer maps; pie chart by populations or by cancers.

2.2.4. European Cancer Observatory

European Cancer Observatory (ECO) is a project developed at the International Agency for Research on Cancer (IARC) in partnership with the European Network of Cancer Registries (ENCR) in the framework of the EUROCOURSE project supported by the European Commission. The ECO platform provides a comprehensive system of information on cancer burden in Europe across three web sites: EUCAN national estimates, EUREG registry data and EUROCIM downloadable data.

EUCAN presents national estimates of cancer incidence, mortality and prevalence for 24 major cancer types in 40 European countries for 2012. The standard methodology used may have produced results different from those developed by national bodies and for appropriate information on population of interest the national information sources should also be used. The **cancer factsheets** show the incidence, mortality and prevalence data for 24 different cancer types in the European Union (27) and in each European country. The following graphics and statistics are available: cancer-specific bar charts; cancer-specific summary tables and cancer-specific interactive maps. The **country factsheets** show the incidence, mortality and prevalence data for 24 different cancer types in the European Union (27) and each of the 40 individual European countries. The following graphics and statistics are available: country-specific bar charts; country-specific summary tables; country-specific pie charts by country for incidence, mortality and prevalence.

EUREG permits the exploration of geographical patterns and temporal trends of incidence, mortality and survival observed in European population-based cancer registries for 35 major cancer entities in about 100 registration areas. It is relatively comprehensive analytic tool and will be here not described in detail.

EUROCIM allows the user to define, extract and request data sets provided by the participating cancer registries.

3. Conclusion

Information sources, which provide data on cancer epidemiology, are nowadays widely accessible in user friendly form with many types of graphical or tabular outputs on national and international level. This is enabled by fast progress of information technologies for data processing, analysis and presentation and by increasing cooperation in the field of cancer epidemiology. Nevertheless, a crucial point of any cancer epidemiological assessment still persists: quality, completeness and representativeness of cancer data. Collection of such data is a complex process requiring close cooperation of wide range of specialists from health care to data management and IT and support of this field is as important as development of cancer epidemiology information tools.

4. References

Cancer Incidence in Five Continents. Section of Cancer Information, International Agency for Research on Cancer, Lyon, France. Available from: http://ci5.iarc.fr/ [cit. 2013-08-05]

CANCERMondial. Section of Cancer Information, International Agency for Research on Cancer, Lyon, France. Available from: http://www-dep.iarc.fr/ [cit. 2013-08-05]

Dušek L, Mužík J, Kubásek M, Koptíková J, Žaloudík J, Vyzula R. Epidemiology of Malignant Tumours in the Czech Republic [online]. Masaryk University, Czech Republic, [2005], [cit. 2013-08-05]. http://www.svod.cz. Version 7.0 [2007], ISSN 1802 – 8861.

European Cancer Observatory.

GLOBOCAN 2008: Estimated cancer Incidence, Mortality, Prevalence and Disability-adjusted life years (DALYs) Worldwide in 2008. International Agency for Research on Cancer, Lyon, France. Available from: http://globocan.iarc.fr. [cit. 2013-08-05]

IHIS CR: Cancer Incidence in the Czech Republic. Prague, Institute of Health Information and Statistics of the Czech Republic, 2013; ISSN: 1210-857X, Available from: http://uzis.cz/en/catalogue/cancer-incidence [cit. 2013-08-05]

WHO Cancer Mortality Database. Section of Cancer Information, International Agency for Research on Cancer, Lyon, France. Available from: http://www-dep.iarc.fr/WHOdb/WHOdb.htm [cit. 2013-08-05]

Introduction to survival analysis

Zdeněk Valenta

Dept. of Medical Informatics & Biostatistics, Institute of Computer Science AS CR, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic

Abstract

Survival analysis is concerned with analyzing time-to-event data where the event of interest usually represents some type of "failure". In clinical medicine, the event of interest may be e.g. death of a patient from well specified causes, autoimmune rejection of the graft by the transplant recipient or other type of graft failure in transplant studies. In certain situations, however, the true survival outcomes may not be observable, because we have observed a so called "censoring event" which prevented the event of interest from occurring. Such censoring event may represent, for instance, loss of a particular subject from follow-up, occurrence of administrative censoring, which typically takes place in clinical trials, or we may indeed observe other type of "failure", e.g. death from fatal injuries rather than from cardiovascular causes which were of primary interest in a particular clinical trial. In this article we will stress the importance of a key assumption relating censoring process to survival outcomes and review principle univariate survival analysis methods for uncorrelated data. We will review popular models for analyzing univariate survival data, many of which enable us quantifying effect the prognostic variables independently exert on survival outcomes. Model examples will cover the classes of non-parametric, parametric and semi-parametric methods. We will also review underlying assumptions of individual models and stress the importance of using appropriate models in analyzing univariate time-to-event data.

Key words

Survival analysis, time-to-event data, censoring process, hazard function, survival time

1. Introduction

In survival analysis we are typically concerned with time to the occurrence of certain type of serious, potentially life-threatening or even terminal medical event, such as patient's organ failure or death, and how this time may be altered using some sort of clinical intervention. We may also study time-to-relapse or to recurrence of the disease where the disease is not terminal. However, in this paper we will limit our focus to non-recurrent uncorrelated univariate events only. Typical examples of the events of interest are, for instance, the occurrence of acute myocardial infarction, stroke, kidney or liver failure, HIV infection, development of AIDS, and, indeed, also death from such causes. The interventions may include various dietary and exercise regimen, but are usually designed to compare the performance of some standard and experimental pharmacological treatment. In clinical settings we study the effect of experimental drugs on the survival of seriously or terminally ill patients, such as in cancer or HIV/AIDS trials. A similar discipline developed in industrial or technical context, where it is called "reliability". One of the first topics of interest in reliability studies was evaluating time to light bulb failure where the exponential distribution was assumed with its memoryless property. In clinical studies time-to-death from certain well-specified cause or disease was historically of primary interest, hence the term "survival analysis" developed for the discipline. Survival analysis is often used in prospective clinical

studies where the excess or rather reduction of the risk under experimental treatment needs to be evaluated relative to some baseline, typically associated with using the placebo or standard treatment. In this paper we will introduce concept of censored data which typically arise in survival studies and will focus on the case of right censoring. We will introduce the notion of hazard and cumulative hazard function, respectively, and show how they are related to estimating the survival function. We will also introduce three kinds of classes for modeling univariate uncorrelated right-censored data, namely the class of non-parametric, parametric and semi-parametric models, respectively. We will also touch upon important assumptions which underline justifiable employment of individual models for analyzing survival data and show some examples of analyzing the data using the R system for statistical analysis and graphics (R Development Core Team, 2013). Some useful monographs and resources dealing with the subject of analyzing right-censored univariate uncorrelated survival data which will not be specifically referred to later in the text include Therneau et al. (2000, 2013), Rosner (1987), Armitage et al. (2008) and Zvárová et al. (2003).

2. Censoring, truncation and related assumptions

In analyzing the survival data we often deal with the fact that we could not observe the value of time to occurrence of the event of interest, e.g. time to death from cardiovascular causes. This frequently occurs in the context of clinical trials which are often designed with a pre-determined end-of-study time point. When the trial is closed we may only conclude that in certain individuals the fatal cardiovascular event did not take place before the study ended, although in fact it could occur soon after that. In this case we speak of so-called "administrative censoring". Censoring can also take place due to patient's withdrawal or because some other event occurred earlier during the study, thus preventing the event of interest from occurring. When, for instance, the patient included in the study died from other than cardiovascular causes before the study closed, than his or her time of death was recorded as a censored observation. It is very important to realize that in survival analysis we are effectively exploiting the information brought about not only by the observed event times, but also the censoring times. In other words, even though until certain time the event of interest was not observed in some individuals and nothing more specific can be said about their future event's occurrence, we are still able to use the former information effectively in survival analysis. Censoring observations thus contribute information which may be effectively utilized in evaluating the survival experience in the respective groups or strata. However, this is only true when a key assumption of the survival analysis is upheld, namely that the censoring process must be independent of the process generating the events under scrutiny. We then speak about "random censoring" or "non-informative censoring". In other words, the reason why for any trial participant the observation time will or will not be censored, may in no way be related to the likelihood of observing a failure for that subject, or, for that matter, to the value of prognostic factors which may influence the survival times in the target population. Speaking of assumptions, one must also make sure that for every trial participant the event may take place only once (i.e. models for recurrent data are not subject of this presentation), entry time into the study must be well defined for each subject and the time scale has to be identical for all subjects enrolled into the trial.

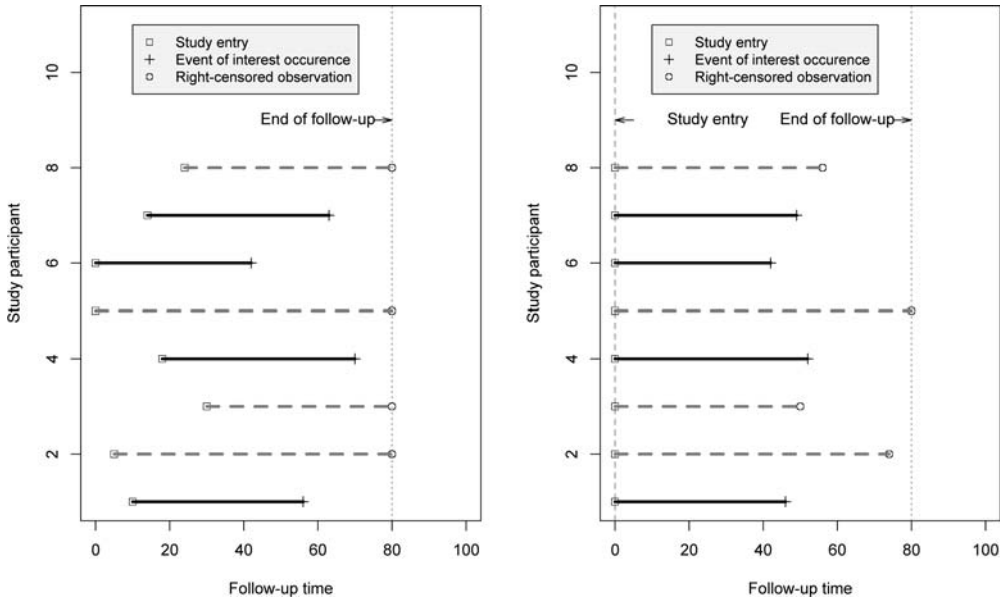


Figure 1. Right-censored survival data description

Let us formally denote time to occurrence of the event of interest X and time to observing a censoring event C. We distinguish between several types of censoring.

- **Right censoring:** Under this scenario we are observing the time $T = \min(X, C)$. Right censoring occurs when $X > C$, otherwise we indeed observe the event of interest occurring at time X. Again, when right censoring takes place, we either did not observe event's occurrence before the study ended, or the patient has withdrawn from the study ("loss from follow-up"), or some other type of event took place thus preventing the event of interest from being observable. Graphical description of right-censored survival data scenario is shown in Figure 1. The right panel reveals that under standard conditions, which are strictly adhered to in clinical trials, the observations' entry into the study may be moved to origin without compromising important assumptions. This is done in order to maximize the number of subjects being at risk at different follow-up times, which further enhances the efficiency of the survival methodology. Note that if we could not move the subjects' entry to origin then, for instance, at time of 2 days into the trial only 2 of 8 available patients would be comprising the risk set.
- **Left censoring:** Left censoring occurs when the event of interest occurred before some well defined point in time, but we are not able to determine exactly when. Let us assume, for instance, that a subject was asked the following question: When did you first time smoke marihuana? If his or her answer was: I smoked pot when I attended secondary school (i.e. before 15 years of age), but cannot tell you exactly when was it the first time it happened, then that is a case of left censoring.
- **Interval censoring:** Under interval censoring scenario we are only able to say that the event of interest occurred at some time within a definite time interval, but are again unable to determine exactly when that happened. This type of censoring typically occurs under periodic patients' follow-up.

- **Truncation:** A different feature in survival studies, sometimes confused with censoring, is called "truncation". For truncated data, only individuals who experience some event are observed by the investigator. The event may be some condition which must occur prior to the event of interest, such as exposure to a disease, entry into retirement center, recurrence of leukemia prior to death, etc. For more details, see e.g. Klein et al, 2003.

3. Hazard rate and survival function

3.1. Hazard rate

Hazard function (or, hazard rate) $\lambda(.)$ is defined as follows:

$$\lambda(x) = \lim_{\Delta x \to 0^+} \frac{P(x \le X < x + \Delta x | X \ge x)}{\Delta x}, \quad (1)$$

where $P(.)$ denotes conditional probability of observing the event of interest within the time interval $[x, x + \Delta(x))$, conditional on the fact that the event did not occur before time x and may only occur at or after that time. Taking the limit while $\Delta(x)$ is approaching zero from the right gives frequency of this conditional probability at infinitesimal time increment after x, thus representing conditional failure rate of the process generating the events of interest. From equation (1) we conclude that product $\lambda(x)\Delta(x)$ approximates conditional probability that the event of interest will take place at infinitesimal time increment after x. This is why in counting process terminology $\lambda(x)$ is referred to as the events' generating process "intensity" (see e.g. Fleming et al, 1991, Andersen et al, 1982).

3.2. Survival and cumulative hazard function

Survival function $S(.)$ gives the probability of not observing the failure in any individual guided by the same events-generating process before or at time x:

$$S(x) = P(X > x) = \int_x^{\infty} f(u)du = 1 - F(x), \quad (2)$$

where $P(.)$ stands for probability measure and $f(.)$ and $F(.)$ are in turn the probability density function (*pdf*) and cumulative density function (*cdf*) associated with the events-generating process. We observe that the survival function is a complement of the *cdf*.

Let us now consider the relationship between the hazard rate and survival function. We will show an intuitive way of uncovering the relationship, which may be more properly shown using the Leibnitz formulae.

$$\begin{aligned} \lambda(x) &= \lim_{\Delta x \to 0^+} \frac{P(x \le X < x + \Delta x | X \ge x)}{\Delta x} = \lim_{\Delta x \to 0^+} \frac{P(x \le X < x + \Delta x)}{\Delta x \cdot P(X \ge x)} \\ &\approx \frac{f(x)\Delta x}{\Delta x \cdot P(X \ge x)} = \frac{f(x)}{S(x)} = -d \ln [S(x)], \end{aligned}$$

hence $S(x) = \exp(-\int_0^x \lambda(u)du) = \exp(-\Lambda(x))$, where $\Lambda(x) = \int_0^x \lambda(u)du$ is called "cumulative hazard rate" (evaluated at time x), or "cumulative intensity" of the process which generated the events, in counting process terminology. It is straightforward that we can formulate the relationship in a reverse manner: $\Lambda(x) = -\ln[S(x)]$.

4. Models for right-censored univariate survival data

4.1. Non-parametric survival models

4.1.1. Kaplan-Meier (Product-Limit) estimator

Standard survival function estimator, called the "Product-Limit estimator" (PL), is attributed to Kaplan and Meier (Kaplan et al, 1958). The PL estimator is given as follows:

$$\hat{S}(x) = \prod_{i:x_i \le x} \left(1 - \frac{d_i}{R_i}\right),$$

where d_i events were observed at time x_i and R_i is the number of individuals at risk at time x_i. The variance of the PL estimator can be estimated using Greenwood's formula:

$$\hat{V}(\hat{S}(x)) = \hat{S}^2(x) \sum_{i:x_i \le x} \frac{d_i}{R_i(R_i - d_i)}.$$

Using the relationship shown above the product-limit estimator can also be used to estimate the cumulative hazard function: $\hat{\Lambda}(x) = -\ln\left[\hat{S}(x)\right]$.

4.1.2. Nelson-Aalen estimator

An alternative estimator of the cumulative hazard function was first proposed by Nelson in a reliability context and lately rediscovered by Aalen who derived the estimator within the counting process framework (Nelson, 1972; Aalen, 1978). It is therefore called "Nelson-Aalen estimator":

$$\tilde{\Lambda}(x) = \sum_{i:x_i \le x} \frac{d_i}{R_i}.$$

Variance of the Nelson-Aalen estimator derived by Aalen is given by:

$$\tilde{V}(\tilde{\Lambda}(x)) = \sum_{i:x_i \le x} \frac{d_i}{R_i^2}.$$

Based on the Nelson-Aalen estimator of the cumulative hazard function, an alternative estimator of the survival function becomes $\tilde{S}(x) = \exp\left(-\tilde{\Lambda}(x)\right)$.

4.1.3. Log rank test

We are often interested in comparing the survival experience in two or more populations of patients, or to test the null hypothesis that the hazard rate in certain population corresponds to a particular rate function. We will describe in more detail "one- sample log-rank test", which is designed to test the null hypothesis H_0 that the hazard rate in the population of interest equals λ_0. This test may then be generalized to compare survival experience in several populations. Using different weight functions we may also obtain different variants of the test.

Now, we wish to test the null hypothesis that population hazard rate equals a particular function, namely H_0: $\lambda(x) = \lambda_0(x)$ for all $x \le \tau$, against the alternative H_1: $\lambda(x) \ne \lambda_0(x)$ for some $x \le \tau$. We will employ the Nelson-Aalen estimator of the cumulative hazard function $\tilde{\Lambda}(x) = \sum_{x_i \le x} \frac{d_i}{R(x_i)}$, where d_i is the number of events observed at event time x_i and $R(x_i)$ is the number of patients at risk just prior to time x_i. The quantity $d_i/R(x_i)$ gives a crude estimate of the hazard rate at an event time x_i. We shall compare the sum of weighted

differences between the observed and expected hazard rates to test the null hypothesis. Let $W(t)$ be a weight function which is zero-valued whenever $R(x)$ is zero. The test statistic for the log-rank test is then given by:

$$Z(\tau) = O(\tau) - E(\tau) = \sum_{i=1}^{D} W(x_i)\frac{d_i}{R(x_i)} - \int_0^{\tau} W(u)\lambda_0(u)du.$$

When the null hypothesis is true, the sample variance of this test statistic is given by

$$V\left[Z(\tau)\right] = \int_0^{\tau} W^2(s)\frac{\lambda_0(u)}{R(u)}du.$$

For large samples, the statistic $Z(\tau)/V(Z(\tau))$ has a central chi-squared distribution when H_0 is true. The most popular choice of the weight function $W(x) = R(x)$ yields a one-sample log-rank test. Other choices lead to Gehan-Wilcoxon, Tarone-Ware, Peto-Peto, modified Peto-Peto, and several variants of the Fleming-Harrington test (see Klein et al., 2003 for more details).

We will now show how R software for statistical computation and graphics (R Development Core Team, 2013) may be used in providing the above mentioned analyses. We will use data on 137 Litoměřice male subjects with the history of acute myocardial infarction between 1991 and 1999. Sixty-eight of them were randomly selected to be intervened on a wide range of cardiovascular risk factors associated with metabolic syndrome, the rest composed a control group. We will compare the survival experience in the two groups of patients using the Kaplan-Meier method and the log-rank test. The event of interest was mortality from acute myocardial infarction (AMI) or ischemic heart disease (IHD), history of these events is recorded in the variable "fail".

```
> summary(survfit(Surv(time,fail) ~ intervention, data=litomerice.muzi.dat, type="kaplan-meier"))
Call: survfit(formula = Surv(time, fail) ~ intervention, data = litomerice.muzi.dat, type = "kaplan-meier")

                intervention=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  326     69       1    0.986  0.0144        0.958        1.000
  707     66       1    0.971  0.0205        0.931        1.000
  823     65       1    0.956  0.0250        0.908        1.000
  852     63       1    0.940  0.0289        0.886        0.999
  894     62       1    0.925  0.0321        0.864        0.991
  915     61       1    0.910  0.0350        0.844        0.981
  936     60       1    0.895  0.0376        0.824        0.972
  964     59       1    0.880  0.0399        0.805        0.962
  985     58       1    0.865  0.0420        0.786        0.951
 1193     56       1    0.849  0.0440        0.767        0.940
 1370     55       1    0.834  0.0458        0.749        0.929
 1697     53       1    0.818  0.0476        0.730        0.917
 2115     52       1    0.802  0.0492        0.711        0.905
 2809     50       1    0.786  0.0508        0.693        0.892
 2842     48       1    0.770  0.0523        0.674        0.879

                intervention=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  683     68       1    0.985  0.0146        0.957        1.000
 1223     66       1    0.970  0.0206        0.931        1.000
 1280     64       1    0.955  0.0253        0.907        1.000
 1416     63       1    0.940  0.0291        0.885        0.999
 2031     62       1    0.925  0.0323        0.864        0.990
 2177     61       1    0.910  0.0352        0.843        0.981
 2328     60       1    0.895  0.0377        0.824        0.972
```

We observe that the survival experience appeared to be a little better in the intervention group. Let us now proceed to formally testing the survival differences in the two groups using the log-rank test. We will also manually verify the value and significance of the test statistic.

```
> (a <- survdiff(Surv(time,fail)~intervention,rho=0,data=litomerice.muzi.dat));
Call: survdiff(formula = Surv(time, fail) ~ intervention, data = litomerice.muzi.dat, rho = 0)

                N Observed Expected (O-E)^2/E (O-E)^2/V
intervention=0 69       15     11.4      1.13      2.58
intervention=1 68        7     10.6      1.22      2.58

 Chisq= 2.6  on 1 degrees of freedom, p= 0.108

 > (a$obs-a$exp)^2/diag(a$var);
[1] 2.58445 2.58445

> 1-pchisq(a$chisq,1);
[1] 0.1079179
```

As may be seen above, a formal log-rank test failed in rejecting the null hypothesis of identical survival experiences (i.e. identical hazard rates) in the two populations of Litoměřice males.

4.2. Parametric survival models

4.2.1. Choices for parametric modeling

Parametric models represent an appealing choice for drawing inference from univariate right-censored survival data. These models are very popular among the researchers because they offer insight into the way the hazard rate changes in time depending on a choice of various model parameters. Here the experience and knowledge of the clinicians together with the information gathered from the actual data may help in selecting suitable model which will best explain the observed data.

In general, a range of continuous probabilistic distributions may be considered for modeling the shape of the hazard rate over time. Some popular distribution choices include the exponential, Weibull, gamma, log-normal, log-logistic, normal, exponential power, Gompertz, inverse Gaussian, Pareto and generalised gamma distribution. For more details regarding the corresponding hazard rate, survival function, probability density function and mean of the distribution the reader is referred to Table 2.2 of Klein et al. (2003), p. 37.

4.2.2. Weibull parametric survival model

Let us now take a closer look at one popular choice of parametric distribution for modeling right-censored univariate survival data, the Weibull distribution. Weibull distribution provides wide flexibility in modeling various trends in the hazard rate, namely modeling the course of the hazard rate as increasing, decreasing or constant.

For a two-parameter (α, η) Weibull distribution formulas for the hazard rate, cumulative hazard and survival function, respectively, are the following: $\lambda(x) = \eta\alpha x^{\alpha-1}$, $\Lambda(x) = \eta x^{\alpha}$, $S(x) = \exp(-\eta x^{\alpha})$.

The log-linear model for $Y = \ln(X)$ may be obtained using the following substitutions: $Y = \ln(X) = \mu + \sigma W$, where W denotes random variable with *extreme values distribution*, and, $\eta = \exp(-\mu/\sigma)$, where $\sigma = 1/\alpha$. In order to include explanatory variable Z into the model, we write: $Y = \ln(X) = \mu + \gamma' Z + \sigma W$.

The Weibull model for the conditional hazard rate $\lambda(x|Z)$ has the following structure:

$$\lambda(x|Z) = (\eta\alpha x^{\alpha-1}) \exp(\beta' Z), \tag{3}$$

where $\alpha = 1/\sigma$, $\eta = \exp(-\mu/\sigma)$, $\beta = -\gamma/\sigma$.

The first component ($\eta\alpha x^{\alpha-1}$) in equation (3) is called "*Weibull baseline hazard*" $\lambda_0(x)$. Note that it is not influenced by the explanatory variables Z and describes hazard level associated with the incidence of the event of interest (e.g. a particular disease incidence or death from

specified causes) in *general population*. This is why this formulation of the Weibull model is called *Weibull proportional hazards model*, because the ratio of the hazard rates at time *x* for two different subjects having distinct covariate values Z_1 and Z_2 only depends on the respective covariate values and regression coefficients β while the baseline hazard cancels out. This phenomenon is described in detail below:

$$\frac{\lambda(x|Z_1)}{\lambda(x|Z_2)} = \frac{\lambda_0(x)\exp(\beta' Z_1)}{\lambda_0(x)\exp(\beta' Z_2)} = \exp\{\beta'(Z_1 - Z_2)\}. \quad (4)$$

Let us now consider fitting the *Weibull parametric model* to our Litoměřice data in R. We are still interested in cardiovascular mortality from AMI or IHD captured in the variable "fail". For this purpose will use the R function *survreg*(.), Please, note that it is important to let the model estimate the actual *scale parameter* value from the data because, unless we wish to specify some specific distributions arising from selecting some particular values. Fixing the scale at 1 or 0.5 will under Weibull distribution lead to fitting the *exponential* and *Rayleigh model*, respectively.

```
> summary(survreg(Surv(time,fail)~intervention+age1,data=litomerice.muzi.dat, dist="weibull",scale=0))

Call: survreg(formula = Surv(time, fail) ~ intervention + age1, data = litomerice.muzi.dat, dist = "weibull", scale = 0)
              Value Std. Error     z        p
(Intercept)  11.5488     1.5755  7.33 2.30e-13
intervention  0.5978     0.3967  1.51 1.32e-01
age1         -0.0495     0.0291 -1.70 8.86e-02
Log(scale)   -0.2369     0.2002 -1.18 2.37e-01

Scale= 0.789

Weibull distribution
Loglik(model)= -230.8   Loglik(intercept only)= -233.4
Chisq= 5.21 on 2 degrees of freedom, p= 0.074
Number of Newton-Raphson Iterations: 8
n= 137
```

Let us recall the relationship between the Weibull parameters α and η, parameters specified for the linear model μ, γ and σ, and the regression coefficients β:

$$\alpha = 1/\sigma,\ \eta = \exp(-\mu/\sigma),\ \beta = -\gamma/\sigma.$$

Because the scale parameter σ was estimated at 0.789, we may obtain the values of regression coefficients β as follows:

$$\beta_{intervention} = -\frac{\gamma_{intervention}}{\sigma} = -\frac{0.5978}{0.789} = -0.7577,$$
$$\beta_{age1} = -\frac{\gamma_{age1}}{\sigma} = -\frac{-0.0495}{0.789} = 0.0627$$

Now it is straightforward to calculate the *hazard ratio* (HR), or equivalently, *relative risk* (RR), associated with both prognostic variables, intervention and age at entry into the trial

$$HR_{intervention} = \exp(\beta_{intervention}) = \exp(-0.7577) = 0.47,$$
$$HR_{age} = \exp(\beta_{age1}) = \exp(0.0627) = 1.06.$$

We observe that there is some indication that after adjusting for the effect of age at entry the intervention of cardiovascular risk factors in male subjects with the history of AMI worked as expected, although the result failed to reach statistical significance at α level 0.05. The hazard reduction in intervened patients is estimated at 53%, each additional year of age appears to increase the hazard by approximately 6%.

4.3. Semi-parametric survival models

4.3.1. Cox proportional hazards model

In one of the most influential seminal papers of all times, D. R. Cox introduced multiplicative model for right-censored univariate survival data (5) involving the baseline hazard component $\lambda_0(x)$ and suggested estimating the regression parameters β using partial likelihood approach (Cox, 1972). This model has a similar form to what we have seen before with Weibull proportional hazards model, only that in this case the baseline hazard function $\lambda_0(x)$ is left completely unspecified, even without the need to be estimated in order to make inference about regression parameters, while in the former case Weibull or some other parametric distribution had to be assumed in modeling the baseline hazard function.

$$\lambda(x|Z) = \lambda_0(x) \cdot \exp(\beta' Z) \tag{5}$$

Let $R(x_i)$, the "*risk set*", denote the number of individuals being at risk of experiencing the event just before time x_i. The probability of any individual failing at time x_i, conditionally on being in the risk set just before time x_i, is given by:

$$P(X = x_i | R(x_i)) = \frac{\exp(\beta' Z_i)}{\sum_{j \in R(x_i)} \exp(\beta' Z_j)}$$

Estimation of regression parameters β proceeds by maximizing *partial likelihood* (6), a concept introduced by D. R. Cox in the above mentioned paper. Partial likelihood allowing for tied observations (due to Breslow) has the following form:

$$PL(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' s_i)}{\left[\sum_{j \in R(x_i)} \exp(\beta' Z_j)\right]^{\delta_i}}, \tag{6}$$

where i indexes the failure times, δ_i gives the number of individuals that failed at time x_i and $s_i = \sum_{j \in \delta_i} Z_j$. Note that only the failure times constitute the terms in the numerator of partial likelihood. The assumption of proportional hazards embedded in the Cox model has exactly the same interpretation as that given in equation (4). Specifically, for any two individuals the hazard ratio at any given time solely depends on the value of their covariates through the regression coefficients, while for the purpose of estimating the model parameters the baseline hazard function does not even need to be specified. However, as is the case with every model assumption, before using the Cox model the proportional hazards assumption needs to be verified for the actual data set at hand.

Lets us now apply the Cox PH model to Litoměřice data and verify the PH assumption for the data. Here we are concerned with the overall survival of the patients for which we will use the variable "scode". The results obtained from fitting the Cox PH model in R are shown below. The first formula gives the summary of regression parameter estimates while the second part performs a formal test of the PH assumption.

```
> summary(cox.fit2 <- coxph(Surv(time,fail)~intervention+age1,data=litomerice.muzi.dat));
Call: coxph(formula = Surv(time, fail) ~ intervention + age1, data = litomerice.muzi.dat)

  n= 137, number of events= 22

                coef exp(coef) se(coef)      z Pr(>|z|)
intervention -0.89698   0.40780  0.47488 -1.889   0.0589 .
age1          0.06221   1.06418  0.03531  1.762   0.0781 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
intervention    0.4078     2.4522    0.1608     1.034
age1            1.0642     0.9397    0.9930     1.140

Concordance= 0.663  (se = 0.063 )
Rsquare= 0.043   (max possible= 0.781 )
Likelihood ratio test= 5.97  on 2 df,   p=0.05056
Wald test            = 5.78  on 2 df,   p=0.05555
Score (logrank) test = 5.93  on 2 df,   p=0.05165

> cox.zph(cox.fit2);
               rho chisq      p
intervention  0.374 2.996 0.0835
age1         -0.145 0.495 0.4818
GLOBAL           NA 3.191 0.2028
```

Age-adjusted intervention effect estimate obtained from fitting the Cox PH model appears to be a little higher than that rendered by the Weibull model. Hazard reduction observed in intervention group is approaching the 60% level. However, the result failed reaching statistical significance (p = 0.0589). The R-function *cox.zph* provides a formal test of proportionality assumption, which appears to be fine for the age at entry while there is a slight suggestion that the intervention effect may not be constant over time (p = 0.0835).

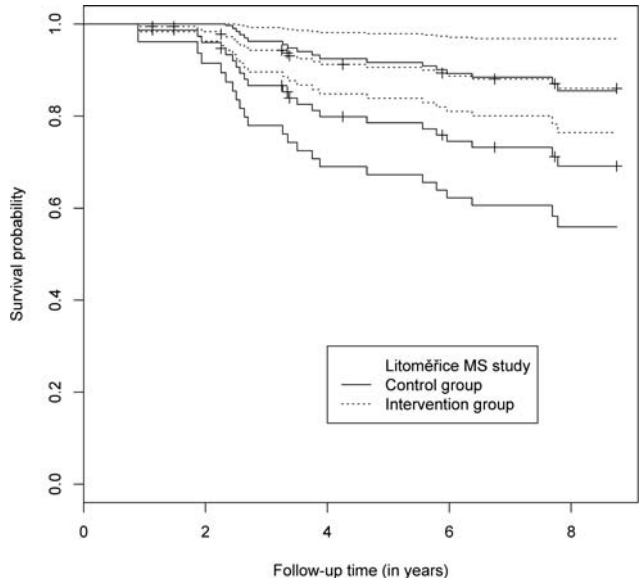


Figure 2. Predicted survival for males aged 55 years based on Cox PH model with 95% confidence limits based on log-transforming the survival function

Let us now proceed to obtaining predicted survival curves (with 95% confidence limits based on the log-transformation of the survival function) for males aged 55 years in the two

follow-up groups based on the fitted Cox PH model. The results are summarized in Figure 2. We observe that the 95% confidence bands corresponding to two groups of males overlap, reflecting the non-significance of the fit obtained earlier from the Cox PH model. Even though not supported with statistical significance of the finding, survival experience appears to be a little more favorable among intervened males aged 55 years than those from the control group.

4.3.2. Gray's time-varying coefficients model

In previous paragraph we observed some indication from our data that the hazard ratio in the two respective groups of males might vary over time. The R-function *cox.zph*, providing a formal test of the PH assumption, rendered the p-value of 0.0835, indicating mildly that the assumption of the hazards proportionality in time may be in question. Let us therefore adopt a different modeling approach proposed by R. J. Gray as an alternative to the Cox PH model (Gray, 1992). Under the penalized modeling framework Gray introduced time-varying regression coefficients (TVC) model allowing for flexible modeling of the hazard ratios over time. The model uses piecewise-constant or cubic B-splines to estimate the values of regression coefficients varying in time. Gray's model may be formally described as follows:

$$\lambda(x|Z) = \lambda_0(x) \cdot \exp(\beta(x)'Z) \tag{7}$$

The R function *coxspline* for fitting the Gray's model is available from Dr. Gray's website at http://biowww.dfci.harvard.edu/~gray/ and may be compiled for all major computational platforms (Unix, Linux, Windows and Mac OS X) in R. It also includes estimator of the survival function for the Gray's piecewise-constant coefficients model proposed and implemented by Valenta (Valenta et al, 2002). Below you may find summary of the results from fitting Gray's TVC model to our data, including a formal test of the PH assumption which differs slightly from that implemented in the *cox.zph* function in R. The column labeled "overall" in the R output below provides results from testing the overall significance of the covariate effects in the Gray's model while the "nonprop" column summarizes results from testing the PH assumption. Note that the latter test found the PH assumption violated for our data (p = 0.0313), thus rendering the Cox PH model unsuitable for summarizing the survival experience in the two groups of our data. The intervention appeared to have significantly reduced cardiovascular mortality overall (p = 0.0317) while the hazard reduction, appearing impressively higher at early stages of the trial, steadily diminished over time.

```
> gm <- cox.spline("t",data$time,fail,spline.cov=x,df=rep(2,dim(x)[2]),nknot=3);
> gm$test;
$intervention
       overall    nonprop
stat 6.0367056 3.17262748
pv   0.0316777 0.03134978
df   1.9995897 1.00021603

$age1
       overall    nonprop
stat 3.0984309 0.09203414
pv   0.1895401 0.95842560
df   2.0002145 1.00038059

> exp(gm$coef)
intervention intervention intervention intervention         age1         age1         age1         age1
   0.2058727    0.2368947    0.6916698    0.8969110    1.0711732    1.0692540    1.0612574    1.0505251
```

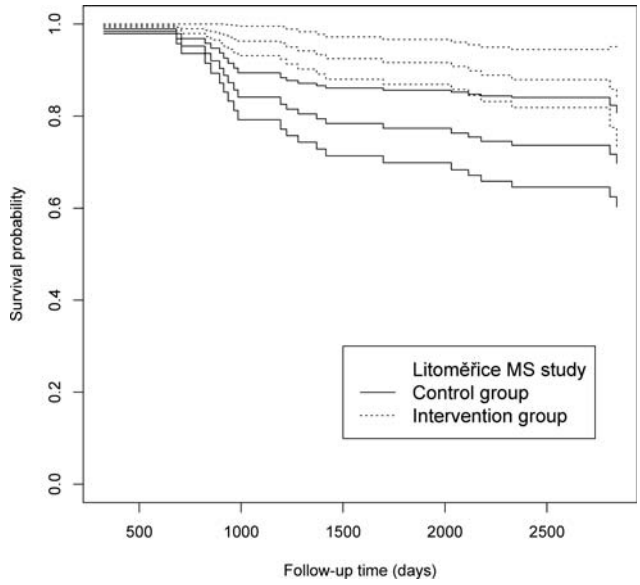


Figure 3. Predicted survival for males aged 55 years based on Gray's TVC model with 95% confidence limits based on log-transforming the survival function

Here we observe that hazard reduction in the intervention group is diminishing over time, initially being close to 80% and finally being reduced to just over 10%. Finally, let us again take a closer look at our findings by estimating the survival functions for males 55 years old. Our findings are summarized in Figure 3. First, we may note that the overall survival experience is indeed much more favorable in the intervention group of males. There is a little overlap of 95% confidence bands after 2 thousand days into the trial, suggesting increasing variability of the regression coefficients as the sample size is reduced later in the trial. Some aspects of choosing an appropriate model for the survival data were also discussed in Valenta et al. (2006).

5. Conclusions

We have reviewed common characteristics and features of uncorrelated univariate survival data, including censoring and truncation. We have provided a deeper insight into the nature of right-censored data and have emphasized important assumption of censoring process being independent from that generating the outcome. We have reviewed classes of non-parametric, parametric and semi-parametric models and have taken a closer look at the principal models representing each class. We have stressed the assumptions of using the statistical models under review and the need for verifying the assumptions for each particular data set to be analyzed.

6. Acknowledgment

The work on this project was partly supported by the long-term strategic development financing of the Institute of Computer Science AS CR (RVO:67985807).

7. References

Aalen OA. 1978. Nonparametric inference for a family of counting processes. Annals of Statistics, 6: 701–726.

Andersen PK, Gill RD. 1982. Cox's regression model for counting processes: A large sample study. The Annals of Statistics, 10: 1100–1120.

Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. John Wiley & Sons, 2008.

Cox DR. 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society - Series B, 20: 187–220.

Fleming TR, Harrington DP. Counting Processes and Survival Analysis, volume 1. John Wiley & Sons, 1991, 2nd edition.

Gray RJ. 1992. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. Journal of the American Statistical Association, 87: 942–951, 1992.

Kaplan EL, Meier P. 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53: 457–481.

Klein JP, Moeschberger MM. Survival Analysis. Techniques for Censored and Truncated Data. New York: Springer-Verlag, 2003.

Nelson W. 1972. Theory and applications of hazard plotting for censored failure data. Technometrics 14: 945–965.

Rosner BA. Fundamentals of Biostatistics. Duxbury, 1987.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

Therneau TM. A Package for Survival Analysis in S, 2013. R package version 2.37-4.

Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer-Verlag, 2000.

Valenta Z, Piťha J, Podrapská I, Poledne R. 2006. Gaining insight from flexible models–assessment of the secondary prevention trial of CHD in the Czech male population with MI history. Methods of Information in Medicine 45:186–190.

Valenta Z, Weissfeld L. 2002. Estimation of the survival function for Gray's piecewise-constant time-varying coefficients model. Statistics in Medicine 21: 717–727.

Zvárová J, Malý M. Statistické Metody v Epidemiologii. Karolinum UK Praha, 2003.

Hazard rate functions driven by finite-state and continuous-state stochastic processes

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno; e-mail: pokora@math.muni.cz

Abstract

The aim of this contribution is to present basic mathematical knowledge on how the hazard rate of the first hitting time is related to the underlaying stochastic process. We would not give complete mathematical treatment, but rather a practical view suitable for modelling and performing simulations.

Keywords

Hazard rate function, first hitting time, multi-state model, continuous-state model.

1. Random process, first hitting time and hazard rate

This contribution is aimed to be used as an introductory text for better understanding Chapter 10 of book (Aalen et al., 2010). It uses similar examples but in more details to show to the reader that, despite the more difficult mathematical background, the application and interpretation of such models is tractable. We hope it could be useful to the readers who are interested in the application of stochastic processes for modelling the hazard rate functions understood as the risk to attain some specific state of the process. It is assumed that the reader already has some knowledge of the basic mathematical notation used in the survival analysis and of its interpretation.

Random process (or stochastic process) $X(t)$ can be understood as a sequence of random variables indexed by instants of time, t. This means, that for each time t, $X(t)$ is a random variable with some probability distribution and statistical characteristics. If the index set is continuous, usually given as interval $[0,\infty)$ or $[0,T]$ for some fixed $T>0$, we call it continuous-time process. If the index set consists only of separated instants of time (finite or countable), the process $X(t)$ is called discrete-time process. Another classification of random processes is by the set of possible values, so called state space. If the state space is a finite or countable set (for example, $\{1,\ldots,5\}$ or $\{1,2,\ldots\}$) the process $X(t)$ is called a chain. If the state space is continuos we say $X(t)$ is a random process with continuous states.

A special class of random processes consists of so called Markov processes (or Markov chains). Markov processes are memoryless random processes. It means, that for fixed instant of time t_0, the future evolution of such a process $X(t)$ depends on the present value $X(t_0)$, not on the history. This mathematical fact is mathematically written in terms of the probability as

$$\mathrm{P}[X(t_0+h)|X(t),0\leq t\leq t_0]=\mathrm{P}[X(t_0+h)|X(t_0)]$$

and says that the probability distribution of the random process $X(t_0+h)$ at every future time t_0+h, conditioned by the knowledge of the past values of the process, is the same as the distribution conditioned only by the knowledge of the present value

$X(t_0)$. Markov processes (and Markov chains) are, in general, the most studied random processes with relatively simple practical application on data or for simulations.

We can think the particular states as different stages of a disease or different therapies. Absorbing state is a state with no further possibility to change the state. If there is no possibility of relapse, the heal of the disease will be represented by an absorbing state. The other states correspond to the survivors. We focus on the hazard rate function, which is an intensity of the survivors at risk of reaching the absorbing state. We derive the dependency of the the hazard rate as on time we examine the so called first hitting time to reach the absorbing state. The first hitting time is a random variable, hence having some probability distribution for which the hazard rate function can be calculated. It will be seen how the initial distribution of the population among the states of the model influences the shape of the hazard rate function.

2. Hazard rate in model with multiple states

Now, we focus on a finite state Markov chain with a single absorbing state. We take a bit simpler example than that shown in (Aalen et al., 2010) and we show how it is possible to obtain different shapes of the hazard rate in this model.

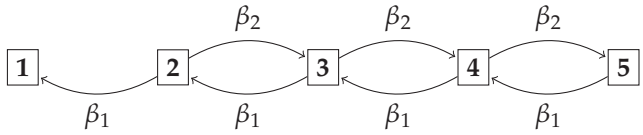


Figure 1. Transition scheme of the model from the example. State 1 is absorbing. Constants β_1 and β_2 are the transition intensities.

Consider the continuos time Markov chain with the state space $1,\ldots,5$. The transition scheme of the specific chain is shown in Fig. 1. Each box represents a state of the chain and the arrows indicate the possible transitions. The parameters β_1, β_2 are the transitions intensities for moving one state down or up, respectively. State 1 is the absorbing state, states 2–4 are states of the survivors. The specific event of our interest is to reach the absorbing state 1. With respect to this event, it could be seen, that the population at risk is concentrated in state 2 only.

Properties of the Markov chain with continuous time can be described by the matrix of transition intensities, which in our example has the following form,

$$\boldsymbol{Q} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \beta_1 & -(\beta_1+\beta_2) & \beta_2 & 0 & 0 \\ 0 & \beta_1 & -(\beta_1+\beta_2) & \beta_2 & 0 \\ 0 & 0 & \beta_1 & -(\beta_1+\beta_2) & \beta_2 \\ 0 & 0 & 0 & \beta_1 & -\beta_1 \end{pmatrix}.$$

The element in row i and column j of $\boldsymbol{Q}$ gives the transition intensity from state i to state j, for $i \neq j$. The elements on the diagonal of $\boldsymbol{Q}$ are calculated in such a way, that the row sums are equal to zero. The zero values of all the row sums of matrix $\boldsymbol{Q}$ is the typical property of the matrix of transition intensities.

The meaning of the transition intensities is better seen when working with a discretized version of the chain. Suppose that the process can evolve only in discrete time steps of given (small) length $\Delta t > 0$. Let us calculate a new matrix

$$\boldsymbol{P} = (\boldsymbol{I} + \boldsymbol{Q})\Delta t,$$

where $\boldsymbol{I}$ stands for the identity matrix of appropriate dimension. If Δt is short enough, all the elements of $\boldsymbol{P}$ have values between 0 and 1, and therefore $\boldsymbol{P}$ can be understood as a matrix of transition probabilities. The value of the element of $\boldsymbol{P}$ in ith row and jth column gives the probability, that an individual in state i at time instant t will move to new state j during the (short) time interval $(t, t + \Delta t]$. Typical property of the transition probability matrix is that all its rows sum to one; such a matrix is called stochastic matrix. Let us choose $\beta_1 = 1$ and $\beta_2 = 1.5$ in our model and let us calculate $\boldsymbol{P}$ for the time interval $\Delta t = 0.02$. We obtain

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.02 & 0.95 & 0.03 & 0 & 0 \\ 0 & 0.02 & 0.95 & 0.03 & 0 \\ 0 & 0 & 0.02 & 0.95 & 0.03 \\ 0 & 0 & 0 & 0.02 & 0.98 \end{pmatrix}$$

We see, e.g., that the probability, that an individual in state 2 will move to state 1 during the time interval of length $\Delta t = 0.02$ is equal to 2 %. Analogously, an individual in state 5 will not change its state with probability 98 %. State 1 is absorbing, which is indicated by the probability of 100 % on the diagonal of the matrix.

The advantage of $\boldsymbol{P}$ is that it gives a simple way to calculate the probability distribution of the population among the states at every multiple of Δt, hence at times $0, \Delta t, 2\Delta t, 3\Delta t, \ldots$. If Δt is chosen short enough, we obtain a pretty good approximation of the original continuous-time Markov chain. Let $\boldsymbol{p}(k\,\Delta t)$ stand for the probability distribution of the population at time instant $k\,\Delta t$. It is a vector consisting of the probabilities that a randomly chosen individual at time $k\,\Delta t$ will be in particular states from 1 to 5. Of course, the probabilities in this vector always sum to one. Similarly, let us denote the initial distribution (at time 0) as $\boldsymbol{p}_0$. Then, we have the general formula

$$\boldsymbol{p}(k\,\Delta t) = \boldsymbol{p}_0\,\boldsymbol{P}^k, \qquad k = 0, 1, 2, \ldots,$$

where $\boldsymbol{P}^k$ stands for the matrix power of order k, or, alternatively, the recurrent formula

$$\boldsymbol{p}[(k+1)\,\Delta t] = \boldsymbol{p}(k\,\Delta t)\,\boldsymbol{P}, \qquad k = 1, 2, \ldots.$$

Now, we can easily perform a simulation of such a process by repeating the calculation according to the last equation for $k = 1, 2, \ldots$. Then, obtained values of the probability distribution of the population among the states can be plotted as functions of time. Resulting distributions obtained from our model are depicted in Fig. 2 for two different settings of the initial distribution of the population. We see that the proportion of the population at state 1 grows in time. The proportions of the population in the other states 2–4 exhibit either decreasing behaviour or it increases at first, reaches a maximum and then decreases as the time grows. The proportions od states 2–4 tend to zero asymptotically, whereas the proportion of state 1 tends to one.

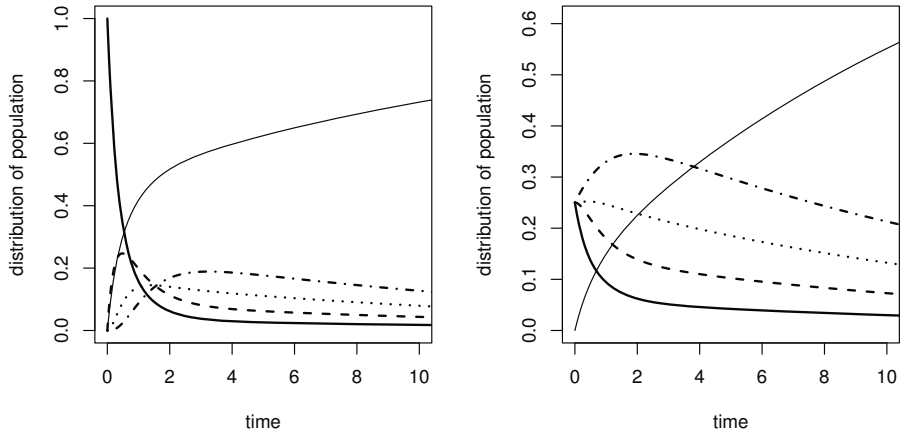


Figure 2. Probability distribution, $\boldsymbol{p}(t)$, of the population in the particular states in time, t: state 1 (thin solid increasing), state 2 (solid), state 3 (dashed), state 4 (dotted) and state 5 (dash-dotted). On the left hand side for initial distribution $\boldsymbol{p}_0 = (0,1,0,0,0)$ (initially in state 2), on the right hand side for uniform initial distribution among the states 2–4, $\boldsymbol{p}_0 = (0, 0.25, 0.25, 0.25, 0.25)$.

Very interesting property of these models is so called quasi-stationary distribution of the population in time. The work quasi-stationary indicates that it is not stationary in general. Only the distribution of the part of the population which survives (hence did not achieve the absorbing state) converges to some stable distribution. To obtain the distribution of the survivors is very easy: we chose only those values of the vector $\boldsymbol{p}(k\,\Delta t)$ which correspond to the survivor (nonabsorbing) states and we normalize it in order to sum to one again. The resulting survivor distribution is denoted by $\boldsymbol{s}(k\,\Delta t)$. In our example, the non-survivors are collected in state 1, hence our $\boldsymbol{s}(k\,\Delta t)$ is obtained by taking only the last 4 values from $\boldsymbol{s}(k\,\Delta t)$ by normalizing the new vector (of length 4). The result is plotted in Fig. 3. We observe the typical behaviour of the distribution of the survivors, it gets stabilized when the time grows. This is not in contradiction with the proportions plotted in Fig. 2: the proportions of the survivors in the whole population tend to zero, but their ratios are kept asymptotically constant. The limiting values of $\boldsymbol{s}(t)$ can be calculated as the normalized elements of the left eigenvector (which is the common eigenvector of the transposed matrix) corresponding to the dominant eigenvalue (i.e. the least absolute eigenvalue) of the submatrix of $\boldsymbol{Q}$ of only the survivor states. In our example, we take the matrix $\boldsymbol{Q}$ without the first row and first column and calculate its eigenvalues: the least one in absolute value is 0.067 and the corresponding left eigenvector is equal to $(0.114, 0.277, 0.504, 0.810)$. By normalization of this vector we get the limiting survivor distribution $\boldsymbol{s}(\infty) = (0.067, 0.163, 0.295, 0.475)$, which are plotted in Fig. 3.

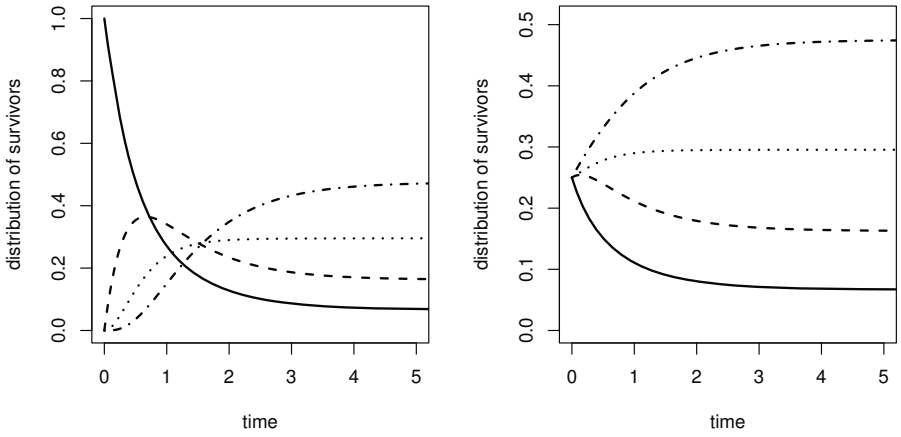


Figure 3. Probability distribution, $\boldsymbol{s}(t)$, of the survivors in time, t: state 2 (solid), state 3 (dashed), state 4 (dotted) and state 5 (dash-dotted). The layout corresponds to the conditions in Fig. 2.

We are interested in the event, when the individuals come to the absorbing state (state 1 in our example). This links the hazard to the distribution of risk for the survivors. If the event of our interest is to pass the individuals to the absorbing state 1, the hazard rate, $\lambda(t)$, at time t is given by summing the probability that an surviving individual is in state j at time t multiplied by the intensity of transition from state j to state 1 for over all the survivor states. This rather complicated formula has simple mathematical notation in form of a scalar product

$$\lambda(k\,\Delta t) = \boldsymbol{s}(k\,\Delta t)'\,\boldsymbol{r},$$

where $\boldsymbol{r}$ is the first column of the transitions intensities matrix $\boldsymbol{Q}$ without the first element; such a vector $\boldsymbol{r}$ gives exactly the intensities of transition from all the survivor states to the non-survivor (absorbing) state 1. In our example, according to its schema and matrices $\boldsymbol{Q}$ and $\boldsymbol{P}$, we see that this can happen only by individuals in state 2 passing to state 1. The resulting hazard rates λ as functions of time are shown in Fig. 4. Note especially the different shapes in accordance with different initial distributions of the population. But, we see that the hazard rate converges to the unique value regardless on the initial distribution of the population. This limiting value can be again easily computed as the absolute value of the dominant eigenvalue of the submatrix of $\boldsymbol{Q}$ corresponding to the survivor states only. In our example, this limiting value is equal to $\lambda(\infty) = 0.067$ and one can check it in Fig. 4, too.

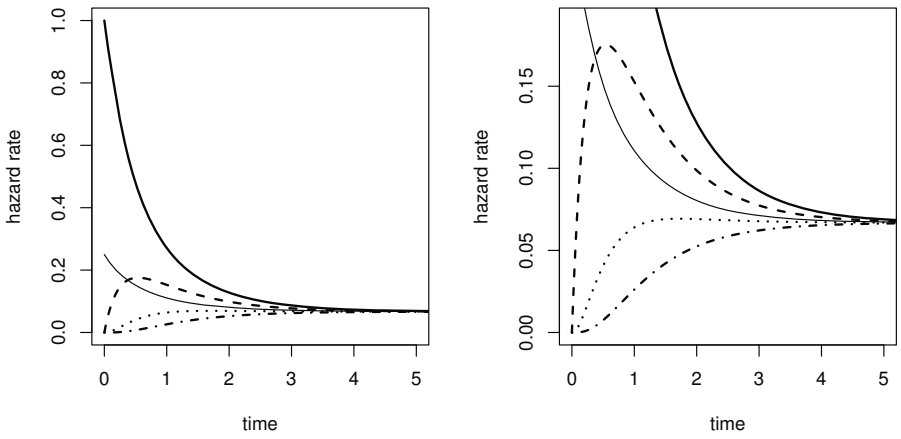


Figure 4. Hazard rates, $\lambda(t)$, (risk to attain the state 1) as functions of time, t, for different initial population distributions: $\boldsymbol{p}_0 = (0, 0.25, 0.25, 0.25, 0.25)$ (thin solid), initially in state 2 (solid), state 3 (dashed), state 4 (dotted) and in state 5 (dash-dotted). The right hand side figure is a detail of the left hand side plot.

3. Hazard rate in model with continuous state space

We continue with the examination of the shape of the hazard rate function driven by a stochastic process. But now, we take the underlaying process with continuos state space. Specifically, let the state space of the process is a positive half line $[0, \infty)$ with an absorbing state 0 which will represent the event of our interest. We skip the usually mentioned random walk and start with a very specific and deeply studied random process called Brownian motion with drift as mentioned in (Aalen et al., 2010). We use this process analogously to describe the evolution of the probability distribution of the population among the state space, which is now assumed to be the interval $[0, \infty)$.

Brownian motion (or Wiener process) with drift is random process $X(t)$ given by the formula

$$X(t) = c - \mu\, t + \sigma\, W(t), \qquad t \geq 0.$$

The parameter $c > 0$ is the initial value of the process, $\mu > 0$ is so called drift parameter and $\sigma > 0$ is a parameter which controls the amount of randomness involved in the process. The variable $W(t)$ is a special random process, which is called standard Wiener process (or standard Brownian motion). Its name is usually referred to Robert Brown due to the similarity of the graph of this process with the trajectory of pollen grains he had observed. This random process $W(t)$ has very special mathematical features, let us mention the most important: for every fixed time $t > 0$, the random variable $W(t)$ has normal distribution with zero mean and variance equal to t, the trajectory (sample path) of $W(t)$ is everywhere continuous but nowhere differentiable. The properties are a bit unusual and in some sence in contradiction with our common thinking. The property of the variance means that the time variable somehow propagates into the value of the process, heuristically written as $[\Delta W(t)]^2 \approx \Delta t$. The nondifferentiability brings causes many problems with the mathematical treatment of the process. Roughly said, when working with Brownian motion (with or without drift) we can not use the common calculus. Instead, so called stochastic differential,

stochastic integral and new rules to deal with such a process must be given. For details, we refer the curious reader to some textbook of stochastic analysis, e.g. (Karatzas and Shreve, 1991).

Nevertheless, the Brownian motion with drift was proved to be an underlaying process for some known shapes of the hazard rate function. And not only in the survival analysis, but also in the theory of reliability, mathematical finance (used for description of the stock prices) in neurophysiology (used for description of the membrane potential). Despite the different mathematical treatment, it is relatively tractable to work with the Brownian motion with drift in simulations. The basic idea uses the property of the normal distribution of the standard Wiener process and is based on the generation of short-time increments, $\Delta X(t)$, from which the resulting process $X(t)$ comes as their cumulative sum,

$$X(0) = c, \qquad X(t+\Delta t) = X(t) - \mu\,\Delta t + \sigma\sqrt{\Delta t}\,\varepsilon(t), \qquad t \geq 0.$$

Here, $\varepsilon(t)$ are elements of a random sample taken from standard normal probability distribution. We choose a time interval of length Δt and generate a large sample $\varepsilon(t)$ from the standard normal distribution. Then the procedure in the last formula is repeated in a loop, until the values of the process $X(t)$ for required time length is obtained. If $Deltat$ is short enough, the result is good approximation of the theoretical Brownian motion with drift.
We generate many, say at least 1000, trajectories with the same parameters. Because we are interested in particular event, entering the absorbing state 0, we calculate so called first hitting time to zero boundary,

$$\tau_0 = \inf\{t \geq 0; X(t) \leq 0\},$$

from each trajectory of $X(t)$. In this way we obtain a sample of the first hitting times to zero boundary and we can plot its histogram and estimate its probability density function. A plot of few trajectories and a histogram of the sample first passage times τ_0 are shown in Fig. 5.

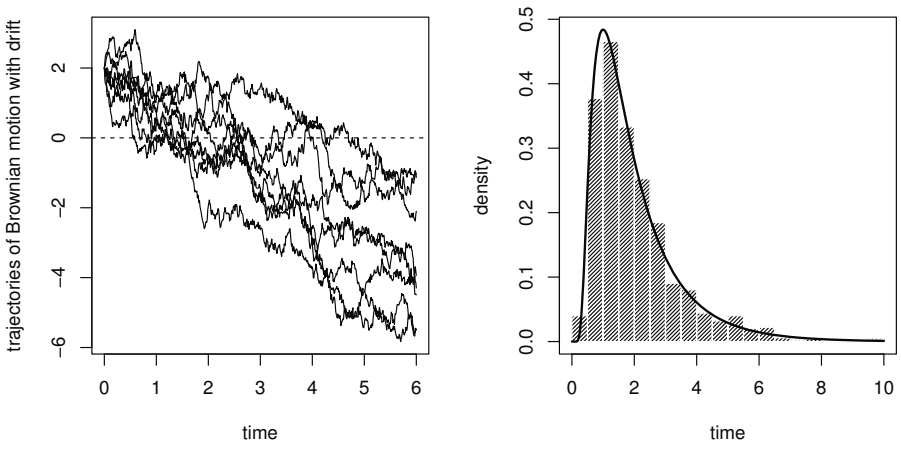


Figure 5. Left hand side: eight trajectories (sample paths) of the Brownian motion with drift. Parameters are $c = 2$, $\mu = 1$ and $\sigma = 1$. Right hand side: histogram of the sample of the first hitting times of 1000 trajectories of the Brownian motion with drift

with the same parameters. The solid curve is the corresponding probability density function of the inverse Gaussian distribution.

It was proved that the first hitting time, τ_0, to the zero absorbing boundary has inverse Gaussian distribution. For comparison, the theoretical probability density function is added to the histogram in Fig. 5. It is given by equation (Chhikara and Folks, 1989; Karatzas and Shreve, 1991)

$$f(t) = \frac{c}{\sqrt{2\,\pi\,\sigma^2\,t^3}} \exp\left[-\frac{(c-\mu t)^2}{2\,\sigma^2\,t}\right].$$

The corresponding survival function is equal to (Chhikara and Folks, 1989)

$$S(t) = \Phi\left(\frac{c-\mu t}{\sqrt{\sigma^2\,t}}\right) - \Phi\left(\frac{-c-\mu t}{\sqrt{\sigma^2\,t}}\right) \exp\left(\frac{2\,\mu\,c}{\sigma^2}\right).$$

Both the functions $f(t)$ and $S(t)$ for some different different initial values c are shown in Fig. 6.

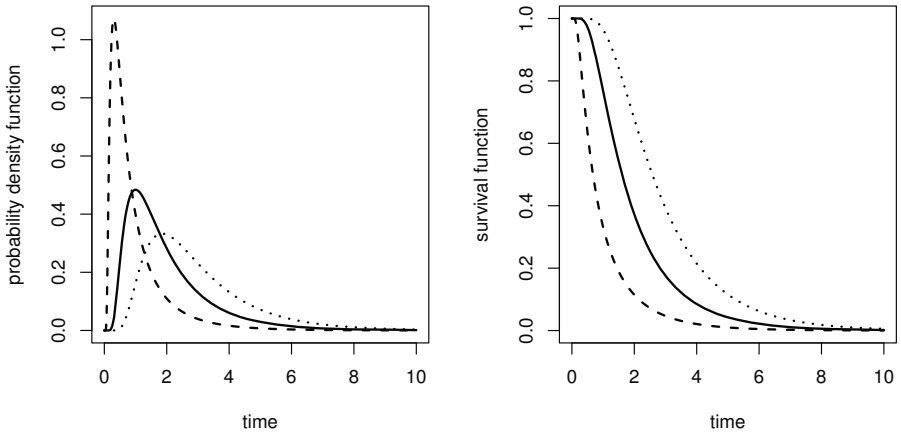


Figure 6. Probability density functions, $f(t)$, (left hand side) and survival functions, $S(t)$, (right hand side) of the inverse Gaussian distribution with initial value $c = 2$ (solid), $c = 1$ (dashed), $c = 3$ (dotted) and parameters $\mu = 1$ and $\sigma = 1$.

The corresponding hazard rate $\lambda(t)$ at time t is equal to $\lambda(t) = f(t)/S(t)$. The shapes of the hazard rate $\lambda(t)$ in dependency on time t are shown in Fig. 7. The hazard rate functions for the Brownian motion with drift exhibit the same stability phenomenon as for the finite-state models. Regardless of the initial value, c, they all converge to the same limiting hazard. If c is close to zero, we get a decreasing hazard rate. For intermediate values of c one gets a hazard which first increases and then decreases; this is the typical shape of the hazard rate for many continuous state space models. For very large large c, an increasing hazard rate is obtained.

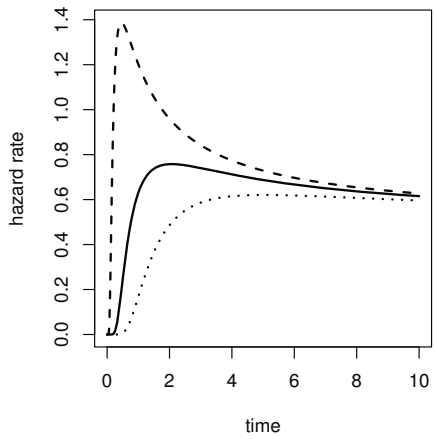


Figure 7. Hazard rate functions, $\lambda(t)$, (risk to attain the state 0) for the same three settings of the initial value and the parameters as in Fig. 5.

Note that the shapes of the hazard rate are very similar to those from the finite state model in the previous section. That is one of the objectives that the continuous state space processes play an important role in the survival analysis. These models are although more difficult to mathematically treat, however, they offer more flexibility to model the shape of the hazard rate, especially when other so called diffusion processes (Brownian motion with drift is one example, another well studied is Ornstein-Uhlenbeck process) are used for the underlaying stochastic process.

At the end of this contribution, we show how simple it is to randomly generate the trajectories of the Brownian motion with drift. We hope it would simplify the way the reader needs to begin to discover the behaviour of the Brownian motion with drift and to obtain the sample of the first hitting times to zero boundary by the simulations. We present the following codes in R language (R Development Core Team, 2012). The first function takes the parameters $c, \mu, \sigma, \Delta t, T$ on its input and returns a vector of the values of $X(t)$ for time instants from 0 to T with the step of length Δt.

```
Bmwd <- function (c, mu, sigma, dt, T) {
# number of increments
n <- ceiling (T / dt)
# generating of random increments
dW <- rnorm (n, mean = 0, sd = 1) * sqrt (dt)
# increments of Brownian motion with drift
dX <- sigma * dW - mu * dt
# cumulative sum if increments (initial=c)
X <- cumsum (c (c, dX))
return (X)
}
```

The second function calculates the first hitting time to zero absorbing boundary from the trajectory of $X(t)$; its parameters are the vector of $X(t)$ and the time step Δt. We note that this leads to a rough approximation of the first hitting time. More precise method for the simulation is given in (Giraudo et al., 2001).

```
Fht <- function (X, dt) {
```

```
# find indices wheres trajectory<=0
hitting.times <- which (X <= 0)
# return NA if there is no such index
if (length (hitting.times) == 0) return (NA)
# else return the first time
return ((hitting.times[1] - 1/2) * dt)
}
```

4. References

Aalen OO, Borgan O, Gjessing HK. Survival and Event History Analysis. Springer, 2010.

Chhikara RS, Folks JL. The Inverse Gaussian Distribution: theory, methodology, and applications. Marcel Dekker, 1989.

Giraudo MT, Sacerdote L, Zucca C. 2001. A Monte Carlo Method for the Simulation of First Passage Times of Diffusion Processes. Methodology And Computing In Applied Probability 3: 215–231.

Karatzas I, Shreve SE. Brownian Motion and Stochastic Calculus. Springer, 1991.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2012.

How to design a parametric survival model

Kateřina Opršalová[1], Jiří Holčík[2]

[1] *Department of Mathematics and Statistics, Faculty of Science, Masaryk University Brno; e-mail: katerina.oprsalova@gmail.com*
[2] *Institute of Biostatistics and Analyses, Masaryk University in Brno; e-mail: holcik@iba.muni.cz*

Abstract

This paper presents some basic principles and methods used when dealing with survival data and developing new results. All processes are illustrated by an example of developing a new approach to parametric regression in survival analysis. Some techniques how to choose the most adequate probability distribution of the data are mentioned. The distributions are transformed in order to get more accurate results. The results are compared with the standard ones according to Akaike‘s information criterion.

Key words

Parametric methods, transformed distribution, survival function

1. Introduction

Every researcher has to face many problems and pass many cross-roads when attempting to obtain new results. The best decision in a single step does not need to be the best for final results. Sometimes it is very tricky to find the simplest and flat way out of the jungle.

This paper can be engaged as a basic guide for those who are starting with their research and are not sure how to get along. It aims at the problem of using parametric methods in survival analysis.

All principles are illustrated by an example of developing a new approach to parametric regression in survival analysis that is an already running author´s research.

1.1. Problem to be solved

Parametric methods are one of the possible tools that can be used to estimate the survival and hazard function for given data. They are not as popular as a Cox model, but they can be useful in some special situations, especially when the assumptions of the Cox model are not met.

The greatest disadvantage of the parametric methods is the need to assume a particular probability distribution of the data. Exponential, Weibull, log-normal and gamma distributions are the most often used for this purpose. All of these distributions are defined in infinite interval $\langle 0; \infty)$, that may theoretically cause overestimation, especially at longer survival times. The above mentioned distributions can respect constant, monotone or unimodal shape of hazard function. Problems arise when it is necessary to model more complex shapes, such as a bathtub shaped hazard function where the hazard declines at the beginning of observed time period, remains almost constant in the middle and rises at the end. Bathtub hazard function occurs very often in survival data. There are some more complex distributions that can be used to handle the problem, such as three parametric

generalized Weibull or generalized gamma distribution. These are not so widely used, because they are not implemented in basic statistical software. Other solution is to take a mixture of more distributions, but the process is not so easy and intuitive.

Our aim is to develop some new distributions defined in a finite interval that will be flexible enough to follow various shapes of hazard function and will avoid the overestimation of the survival function in longer survival times.

1.2. Data

All of the used processes should be optimized for particular data. We use data of 333 women who suffered from breast cancer diagnosed in stage IV of the disease in Czech Republic in 1990. The analysed sample contains 8 right censored observations. Maximum observed uncensored survival time is 5753 days.

It can be useful to plot a graph of nonparametric estimates of the survival and hazard functions to see the basic character of the data.

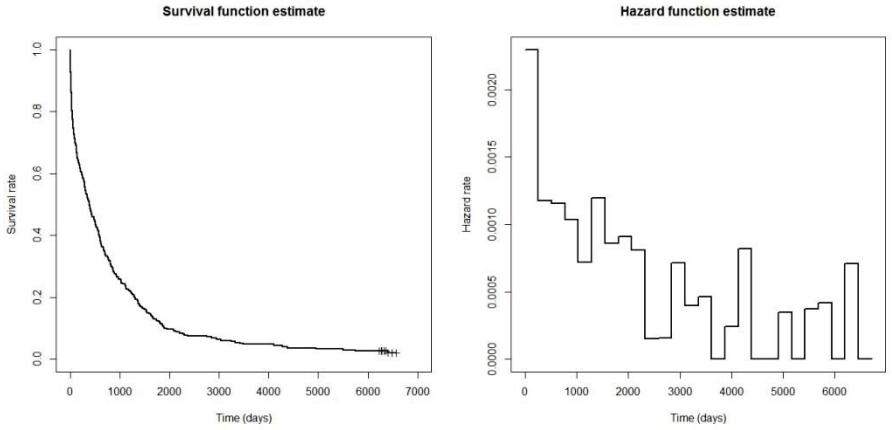


Figure 1. Kaplan-Meier nonparametric estimate of survival function and nonparametric bin estimate of hazard function

As we can see, all of the censored survival times are situated at the end of the studied time interval. Probability of survival slumps to the time of 2000 days and after this time declines slowly. The hazard of death declines, but slight growth can be seen at the end of the observed time interval.

2. Methods

The step-by-step process of the research will be described in this section. We start with parametric distribution selection and the estimation of its parameters, which is the basic proceeding in parametric modelling. We determine our own problem when estimating survival function of the given data and propose its solution. We work with functions without covariates only to show basic principles. All results can be extended to deal with covariates.

2.1. How to choose the most appropriate distribution of the data

The first step when using parametric methods is choosing the probability distribution that is the most appropriate for given data. There are many methods how to perform this selection.

One option is using various distributions to estimate the survival function without the data character consideration. This can be very long and computationally tedious process.

Other way is to think about the data and try to find the distribution that is the most suitable for its character and construction. Shape of the hazard function can be one of the selection criterions. Exponential distribution, as the simplest one, can be used to model constant hazard. Weibull and Gamma distributions can respect monotone course. Log-normal distribution is suitable for modelling the hazard curves with one vertex.

We can inspire ourselves with older researches that worked with the similar data and use the same distribution as mentioned there. This approach, however, is not actually proper in many situations. Every small variation in data characteristics or research conditions can cause indispensable changes in final result.

We can also use some numerical methods for distribution fitting. Unfortunately, these methods can be computationally demanding and require wide analytical capabilities. The problems with the additional accuracy of numerical methods often arise.

Usually, researcher tries to use the most simple and effective way to reach the goal. Graphical methods are the best choice for the purpose. They are used for displaying and interpretation of the data. The basic idea of graphical methods used in distribution fitting is to see whether the survival time, or a function of it, has linear relationship with the distribution function and the cumulative hazard function of a given parametric distribution, or functions of them. The graphical demonstration of such a relationship should be a straight line. The distribution that subjectively fits the straight line most precisely should be the best choice for us.

Two most popular graphical methods are probability and hazard plotting. The basic idea of these methods is estimating the sample cumulative distribution function (or cumulative hazard function) and its comparison with a selected theoretical distribution for the survival time (Lee and Wang, 2003). The principal difference between the two approaches is that the hazard plotting is designed to handle censored data.

We can try to fit many various distributions to see which one gives the best results for our data.

After applying the standardly used distributions mentioned in paragraph 1.1, probability and hazard plotting demonstrate that the Weibull distribution seems to be the most appropriate for our data.

2.2. How to estimate distribution parameters

Although many methods for estimating parameters have been developed, only a few of them are able to work with censored data.

Maximum likelihood method is the most widely used one. We have to change the standard maximum likelihood method, so that it would be able to handle censored data. It acts as standard maximum likelihood for uncensored data and replaces probability density function by survival function when censoring occurs (Hosmer and Lemeshow 1998).

We can work with a likelihood function in form

$$L\big((\boldsymbol{t},\boldsymbol{c}),\boldsymbol{\beta}\big) = \prod_{i=1}^{n}\{[f(t_i,\boldsymbol{\beta})]^{c_i}[S(t_i,\boldsymbol{\beta})]^{1-c_i}\}, \quad (1)$$

or apply its logarithmic form shaped as

$$l\big((\boldsymbol{t},\boldsymbol{c}),\boldsymbol{\beta}\big) = \sum_{i=1}^{n}\{c_i\, ln[f(t_i,\boldsymbol{\beta})] + (1-c_i)\, ln\, S(t_i,\boldsymbol{\beta})\}, \quad (2)$$

where $f(t_i,\boldsymbol{\beta})$ is a probability density function and $S(t_i,\boldsymbol{\beta})$ a survival function for a survival time t_i, $\boldsymbol{\beta}$ is a vector of estimated parameters and $\boldsymbol{c}$ is an indicator of censoring defined as

$$c_i = \begin{cases} 1, & \text{if } t_i \text{ observed;} \\ 0, & \text{if } t_i \text{ censored.} \end{cases}$$

After obtaining the likelihood (or log-likelihood) function, partial derivatives with respect to parameters from vector $\boldsymbol{\beta}$ are taken and the likelihood equations are determined in form

$$\frac{\partial l\big((\boldsymbol{t},\boldsymbol{c}),\boldsymbol{\beta}\big)}{\partial \beta_j} = 0\,. \quad (3)$$

We obtain maximum likelihood estimators of parameters solving the system of the above mentioned equations. Usually, the explicit solution cannot be found. In such a case the numerical optimization has to be used. It brings some other problems with computation, such as initial values selection.

As to particular results, we obtain log-likelihood function for Weibull distribution in shape

$$l\big((\boldsymbol{t},\boldsymbol{c}),(\lambda,b)\big) = \sum_{i=1}^{n}\left\{c_i \ln\left(\frac{b}{\lambda^b}\, t_i^{b-1} e^{-\left(\frac{t_i}{\lambda}\right)^b}\right) + (1-c_i)\ln\left(e^{-\left(\frac{t_i}{\lambda}\right)^b}\right)\right\}. \quad (4)$$

After maximizing this function we get the parameter estimators

$$\hat{\lambda} = 620.270 \quad (5)$$
$$\hat{b} = 0.617$$

We can use the Weibull distribution with these estimated parameters to fit the survival function for our data and compare the result with the Kaplan-Meier estimator of survival function shown above.

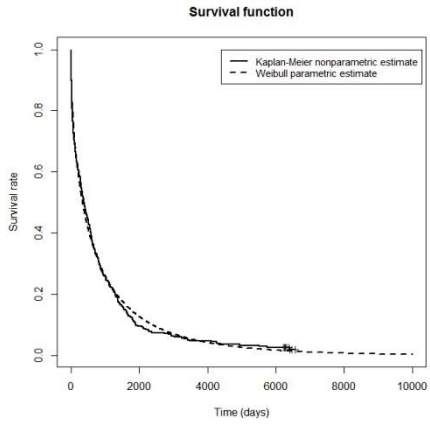


Figure 2. Comparison of nonparametric Kaplan-Meier and parametric Weibull survival function estimates

2.3. How to improve the estimates

Previous picture shows that Weibull distribution fits the survival function quite precisely. However, it overestimates the probability of surviving around 2000 days and underestimates it in longer observed survival times. Problem can come out when extrapolating the results into the future. Weibull distribution is defined in the interval $\langle 0,\infty)$, thus there exists some probability, that the patient can live, for example, 1000 years. This is of course not possible.

It can be useful to find some other distribution that is defined in finite interval. This means there exist some end points that can be interpreted as a maximum possible survival time for a patient with particular diagnosis.

We can try to transform the standardly used distributions to follow the above mentioned properties.

2.3.1. What type of transformation function to choose?

We need to find some transformation function f with values in range $\langle 0,a)$, where a is the maximum value. This means that $f(t) = 0$ for $t = 0$ and $f(t)$ asymptotically approaches a as $t \to \infty$.

As examples of such functions we can introduce

$$y_1(t) = \frac{akt}{1+kt}$$
$$y_2(t) = a(1 - e^{-kt})$$
$$y_3(t) = ae^{\left(-\frac{k}{t}\right)} \qquad (6)$$
$$y_4(t) = \frac{2a}{1+e^{-kt}} - a$$
$$y_5(t) = \frac{2a}{1+(1+\varepsilon)^{-kt}} - a$$

where $k > 0$ is the shape parameter.

But which of the functions is the most appropriate? We can determine some additional characteristics of the functions. If it is not possible to do so, the only chance is to apply more functions and choose that with the best results. The additional properties can arise during the process.

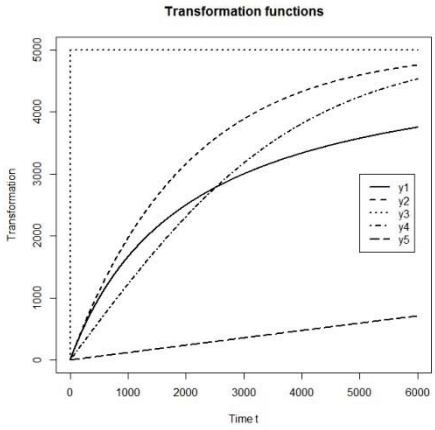


Figure 3. Shapes of various transformation functions with given values of parameters $a = 5000$, $k = 0.0005$, $\varepsilon = 0.1$

2.3.2. How to transform the standard distributions

Suppose that $F(t)$ is a cumulative distribution function and $f(t)$ is the probability density function of the standard distribution. $G(t)$is a cumulative distribution function and $g(t)$ is the probability density function of the newly obtained distribution. y is a transformation function.

Basic properties of the transformed distribution can be taken using following formulas:

$$G(t) = P(T \le t) = P(y \le t) = P(T \le y^{-1}) = F(y^{-1})$$

$$g(t) = G'(t) = f(y^{-1}) \cdot (y^{-1})' \tag{7}$$

$$S(t) = 1 - G(t)$$

$$h(t) = \frac{g(t)}{S(t)}$$

The results can be compared according to the value of log-likelihood function obtained when estimating the parameters of transformed distributions.

After transforming the Weibull distribution using all five transformation functions and applying the maximum likelihood method we obtain following values of log-likelihood function.

Table 1. Values of log-likelihood for different transformation functions

Transformation function	Log-likelihood
y_1	-2453.862
y_2	-2460.994
y_3	-2769.675
y_4	-2450.197
y_5	-2453.192

We obtain the highest value of likelihood function for the transformation function $y_4(t) = \frac{2a}{1+e^{-kt}} - a$ which is part of the sigmoidal function.

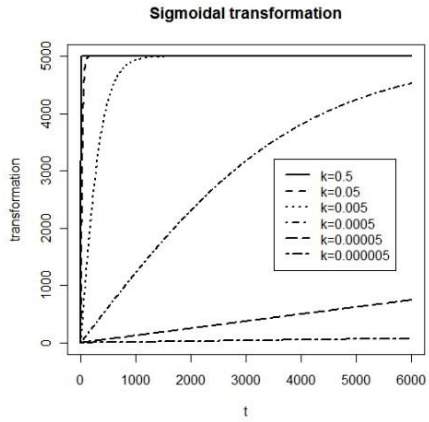


Figure 4. Sigmoidal transformation for various values of k

3. Results

Weibull distribution seems to be the most appropriate to fit the breast cancer data according to probability plot. The distribution is transformed by sigmoidal transformation function. Basic descriptive functions of the newly obtained transformed Weibull distribution are as follows:

Probability density function

$$f(t) = \frac{2ab}{l(a^2-t^2)}\left(-\frac{1}{l}ln\frac{a-t}{a+t}\right)^{b-1} e^{-\left(-\frac{1}{l}ln\frac{a-t}{a+t}\right)^b} \quad (8)$$

Cumulative distribution function

$$F(t) = 1 - e^{-\left(-\frac{1}{l}ln\frac{a-t}{a+t}\right)^b} . \quad (9)$$

Survival function

$$S(t) = 1 - F(t) = e^{-\left(-\frac{1}{l} ln \frac{a-t}{a+t}\right)^b} .(10)$$

Hazard function

$$h(t) = \frac{f(t)}{S(t)} = \frac{2ab}{l(a^2 - t^2)} \left(-\frac{1}{l} ln \frac{a-t}{a+t}\right)^{b-1} . \quad (11)$$

Note that all of the equations contain parameter l that is a component neither of the Weibull distribution, nor of the transformation function. In fact, it is a product of the distribution parameter λ and the parameter k that comes from transformation function. These two parameters occur always in product, so they can be replaced by parameter $l = \lambda k > 0$.

The next step is estimation of the parameters by the maximum likelihood method. It is applied in the form of

$$l\big((\boldsymbol{t}, \boldsymbol{c}), (a, b, l)\big) = \sum_{i=1}^{n} \left\{ c_i \ln \left(\frac{2ab}{l(a^2 - t_i^2)} \left(-\frac{1}{l} \ln \frac{a-t_i}{a+t_i}\right)^{b-1} e^{-\left(-\frac{1}{l} \ln \frac{a-t_i}{a+t_i}\right)^b} \right) + (1 - c_i) \ln \left(e^{-\left(-\frac{1}{l} ln \frac{a-t_i}{a+t_i}\right)^b} \right) \right\} (12)$$

Maximum likelihood estimates of the parameters obtained for the experimental data are

$\hat{b} = 0.613$

$\hat{l} = 0.148$ (13)

$\hat{a} = 8357$ days

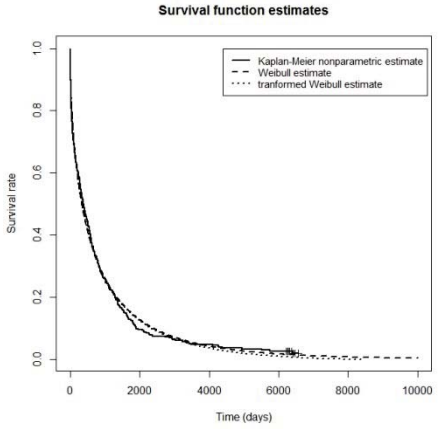


Figure 5. Estimated survival functions using Weibull and transformed Weibull distributions

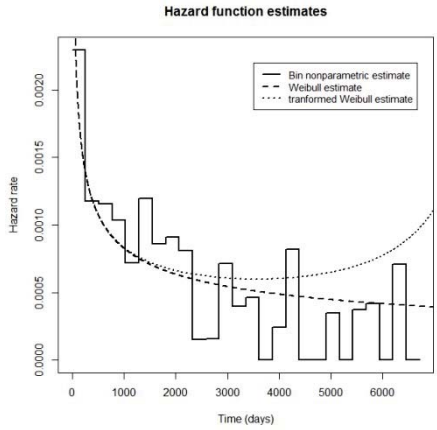


Figure 6. Estimated hazard functions using Weibull and transformed Weibull distributions

4. Conclusions

Transformed Weibull distribution seems to fit the survival function similarly to Weibull distribution. Small differences arise for the longer survival times observed. Contrary to the standard Weibull distribution that goes to infinity, the transformed Weibull distribution ends at time 8357, which is about 23 years. This survival time is reasonable and can be considered as estimated maximum possible survival time.

We can use Akaike´s information criterion as a formal comparative tool (Bradburn et al. 2003). The Weibull distribution is more appropriate than the transformed Weibull distribution according to the criterion, but the difference is only about 2‰.

It is mostly caused by greater complexity and higher number of parameters of the transformed Weibull distribution. Nevertheless, the distribution has some special properties that can make it more suitable in some situations. Its great benefit is the ability to fit the bathtub hazard function. Standard Weibull distribution can fit only a monotone one. The other gain is the already mentioned estimate of maximum survival time.

5. References

Bradburn M J, Clark T G, Love S B, Altman D G. 2003. Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. British Journal of Cancer 89: 605-611.

Hosmer DW, Lemeshow S. Applied Survival Analysis: Regression Modeling of Time to Event Data.NY: John Wiley&Sons, 1998. 386 p. ISBN 978-04-711-5410-5.

Lee E T, Wang J W. Statistical Methods for Survival Data Analysis. NY: John Wiley&Sons, 2003. 534 p. ISBN 978-04-713-6997-4.

Sample size and power analysis in epidemiological and clinical research

Pavla Kadlecová

ADDS s.r.o, Brno; e-mail: pkadlecova@adds.com

Abstract

Determination of appropriate number of subjects to be included in an epidemiological or clinical study is one of the most important tasks in designing of the study. The number of subjects enrolled has direct relation to probability of true significant results (statistical power), to duration and to costs of the study. The sample size is estimated in order to achieve sufficient power which depends on expected difference between compared groups, type of analyzed variable (binary, continuous, censored), type of hypothesis (superiority, non-inferiority, equivalence), and other factors, e.g., variability of data. Further, understanding of factors which affect the power of statistical analyses is also important in statistical considerations and interpretations of statistical results. In this article the main aspects of sample size calculation and power analysis are explained and practical examples are presented.

Key words

Sample size, power, clinical research

1. Introduction

Sample size, power of statistical test (i.e. probability of rejection of H_0 if it is really false) and difference between compared groups (e.g. treatment effect in placebo-controlled trial) are three closely related milestones of inferential statistical analysis.

It is obvious that sample size estimation is particular in designing of studies which compare two groups; however, the sample size should be justified also in case of observational studies in order to achieve sufficient precision of characteristics' estimates.

The sample size/ power calculation and consideration is critical in designing of the clinical trial; however, is also important in interpretation and justification of statistical results, especially in lack of significant result.

2. Basic statistical consideration

2.1. Type I, Type II error, alpha and beta

Rejection of H_0 if the hypothesis is true is called type I error. Probability of type I error is called alpha (level of significance).

Not rejecting of H_0 if the hypothesis is false is called type II error. Probability of type II error is called beta.

Power is defined as 1-beta, i.e. probability of rejection of H_0 if the hypothesis is false.

Table 1. Type I and II error

	Reality	
Results of test	**H_0 is true**	**H_0 is false**
H_0 not rejected	Correct conclusion	Type II error (beta)
H_0 rejected	Type I error (alpha)	Correct conclusion (1-beta) = power

2.2. Three communicating vessels (effect – power – sample size)

The most common task of sample size calculation/ power analysis is related to comparison of two groups, i.e. testing of the following hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{VS.} \quad \begin{array}{l} H_1 : \theta \neq \theta_0 \\ H_1 : \theta < \theta_0 \\ H_1 : \theta > \theta_0 \end{array}$$

The sample size is mainly affected by the following factors:

1. Level of significance
2. Power
3. Effect (difference in compared groups – treatment groups, males vs. females, with vs. without disease,…)

Level of significance

Sample size is highly dependent to the level of significance; however, the level of significance usually given by guidelines. Standardly used level of significance is 0.05. In some cases level of significance 0.01 could be required. In case of multiple testing the level of significance needs to be adjusted by appropriate manner (e.g. using by Bonferroni correction).

Effect, power and sample size

Difference between the groups (effect), power of statistical test and sample size are very closely related. First, the higher sample size gives us higher power to reject the H_0, i.e. we will have higher probability to demonstrate the difference between groups if it really exists (Figure 1). Second, higher effect gives us higher power to reject H_0, therefore, lower sample size is necessary (Figure 2, Figure 3).

With very high sample size we will be able to demonstrate significant difference between groups although size of the difference could be very small. Important is to demonstrate meaningful (e.g. clinically relevant) difference as statistically significant, especially for non-inferiority and equivalence studies the determination of margin is one of the critical point during designing of the study (see section 3).

Sample size of clinical trials is usually calculated to achieve power 80% or 90%.

The relationship effect-power-sample size needs to keep on mind also during analysis and interpretation of statistical results. In epidemiological research, the large sample sizes are not exceptional. In order to make appropriate interpretation of the results the size of the demonstrated effect needs to be taken account. Demonstration "only" statistical significance is not sufficient. In case of lack of power to demonstrate statistical significance it is useful to perform power analysis on given data in order to see which power was achieved and how many subjects would be needed to achieve e.g. 80% power.

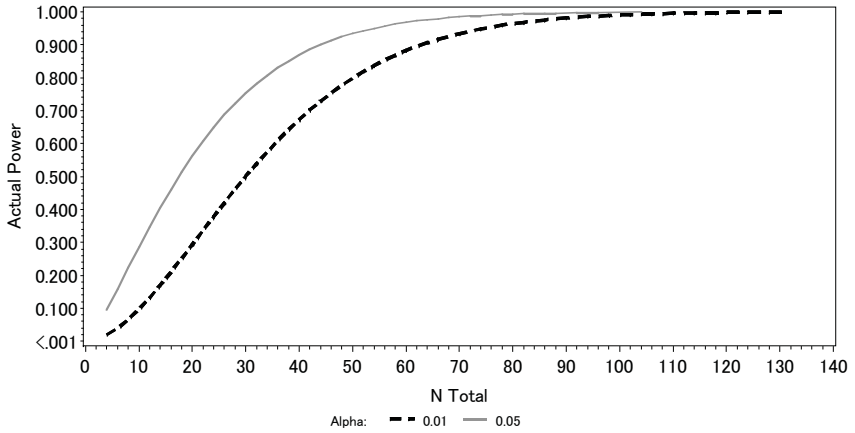


Figure 1. Relationship between sample size and power of t-test for fixed effect

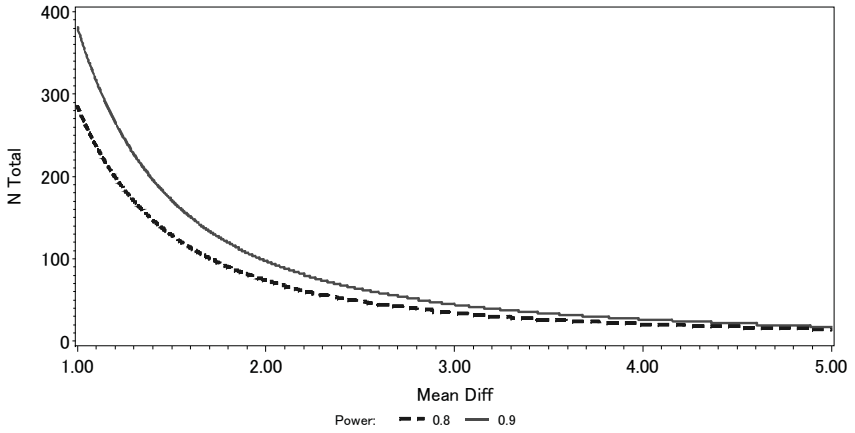


Figure 2. Relationship between effect and sample size for fixed power 80% and 90% of t-test

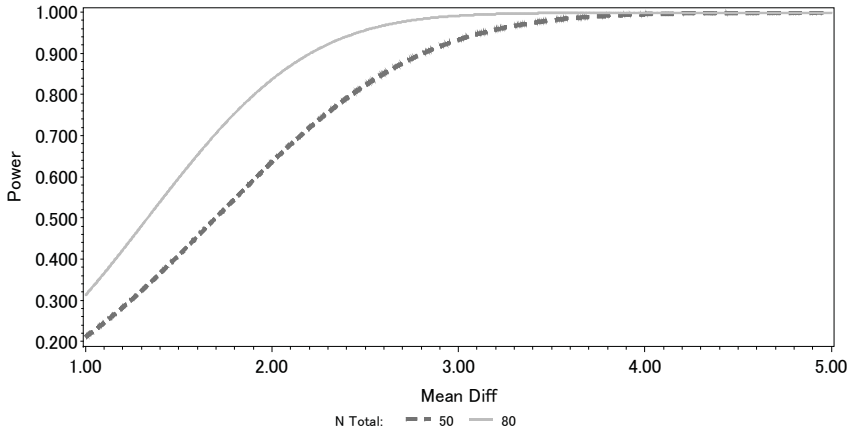


Figure 3. Relationship between effect and power t-test for fixed sample size 50 and 80 subjects

3. Role of sample size/ power calculation in clinical trials and in interpretation of results of statistical analyses

Sample size/ power calculation in designing of clinical trials

The number of subjects/patients to be enrolled in the clinical trial needs to be fixed before the start of the study. Thus, the sample size needs to be determined in order to enable sufficient power to achieve objective of the trial with given design. In general the designing of the clinical trial consist of the following steps:

1. Determination of "clinical" hypothesis, i.e. what we would like to show, demonstrate, determine in our study
2. Selection of primary endpoint (clinical parameter on continuous scale like blood pressure, level of glycosylated hemoglobin, incidence of disease/ cardiovascular event, relapse of disease, quality of life)
3. Selection of design (parallel, cross-over)
4. Specification of the statistical hypothesis, methods of analysis
5. Assumptions based on results previous studies, review of articles, analysis of epidemiological data
6. Sample size calculation (drop-out needs to be add up)
7. Consideration whether it is possible to enroll such a number of subjects regarding time, cost, incidence of disease (in given region) usually follows.

Role of sample size/ power consideration in interpretation and justification of analysis results

At least power consideration should be a part of interpretation and justification of the results of analysis, especially in lack of significant result. The following questions should be answered in order to interpret the results appropriately:

- In case of significant results, has the difference (or another statistic odds ratio or hazard ratio) meaningful size?
- Was the lack of significant results caused by small difference between compared groups?
- Was the lack of significant results caused by small difference between compared groups?
- Or did not we have sufficient sample size? If not, how many patients we would need.
- What about the previous research – how many patients they had to demonstrate the significance. Was the same primary parameter (binary, continuous) used in previous research? Was the statistical method used in previous research the same? There is some reason why our results should differ?
- Was there another factor which influences the results? If possible we should adjust the results for that factor.
- Is not high variability in data caused by difference characteristics of the subjects? If yes, it usually leads to adjustment or subgroup analysis.

4. Common task for sample size calculation

4.1. Superiority, non-inferential, equivalence study

Superiority trial is designed to demonstrate difference between treatments.

The tested hypothesis is: H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$, where θ is parameter of tested group and θ_0 is parameter of reference group (Figure 4). The input for sample size/ power calculation is the null hypothesis, i.e. that we will test that the difference is zero (as the most often case), assumed size of effect and further factors discussed in section5.

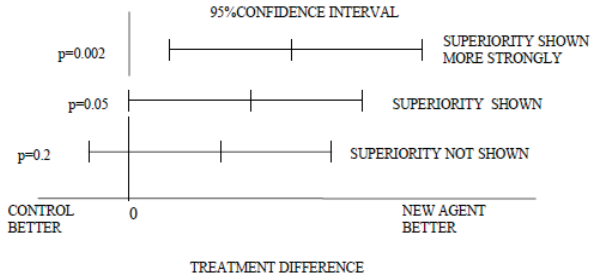


Figure 4. Relationship between significance test of superiority trial and confidence interval. Source: CPMP/EWP/482/99

Non-inferiority trial is designed to demonstrate that the new treatment has not less effect, i.e. that is more effective or have the same effect as reference treatment. Non-inferiority margin has to be a priory defined.

The tested hypothesis is: H_0: $\Delta_N < -\Delta$ vs. H_1: $\Delta_N > -\Delta$ where Δ_N is detectable difference between tested and reference groups (tested - reference) and Δ is non-inferiority margin (Figure 5). The input for sample size/ power calculation is the null hypothesis, i.e. that we

will test that the difference is above -Δ, the size of Δ (usually based on guidelines or clinical consideration), assumed size of effect and further factors discussed in section 5. The sample size needed in non-inferiority trial is usually higher than for superiority (depending on assumptions and the margin).

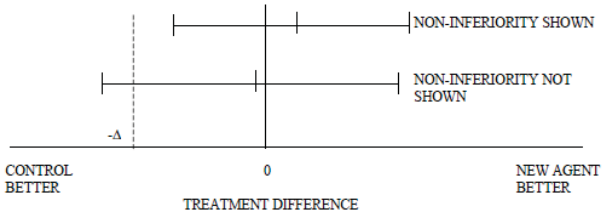


Figure 5. Relationship between significance test of non-inferiority trial and confidence interval. Source: CPMP/EWP/482/99

Equivalence trial is designed to confirm absence of meaningful difference between tested and reference treatment (e.g. used to confirm that both tested and another treatment has equivalent distribution of the active substance in a body). Equivalence margin has to be a priory defined.

The tested hypothesis is: H_0: $\Delta_E < -\Delta$ OR $\Delta_E > +\Delta$ vs. H_1: $-\Delta < \Delta_E < +\Delta$, where Δ_E is difference between tested and reference treatment (tested - reference) parameter of tested group and Δ is equivalence margin (Figure 6). The input for sample size/ power calculation is the null hypothesis, i.e. that we will test two one-sided hypothesis with the margin Δ, the size of Δ (usually based on guidelines, e.g. 80-125%, or clinical consideration), assumed size of effect and further factors discussed in section5.

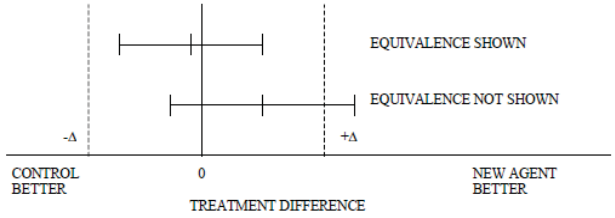


Figure 6. Relationship between significance test of equivalence trial and confidence interval. Source: CPMP/EWP/482/99

4.2. Phase IV study and precision estimates

The current practice is that sample size for Phase IV studies has to be justified. The Phase IV studies are "post-marketing surveillance (PMS) studies but every PMS study is a phase IV study. Phase IV is also an important phase of drug development. In particular, the real world effectiveness of a drug as evaluated in an observational, non-interventional trial in a naturalistic setting which complements the efficacy data that emanates from a pre-marketing randomized controlled trial (RCT)."(Suvarna 2010)

Therefore, the objective of the study is usually to determine the effect of the treatment in clinical practice. The sample size justification is based on precision estimate, usually width

of confidence interval. In order words the sample size is determined to provide sufficiently narrow confidence interval.

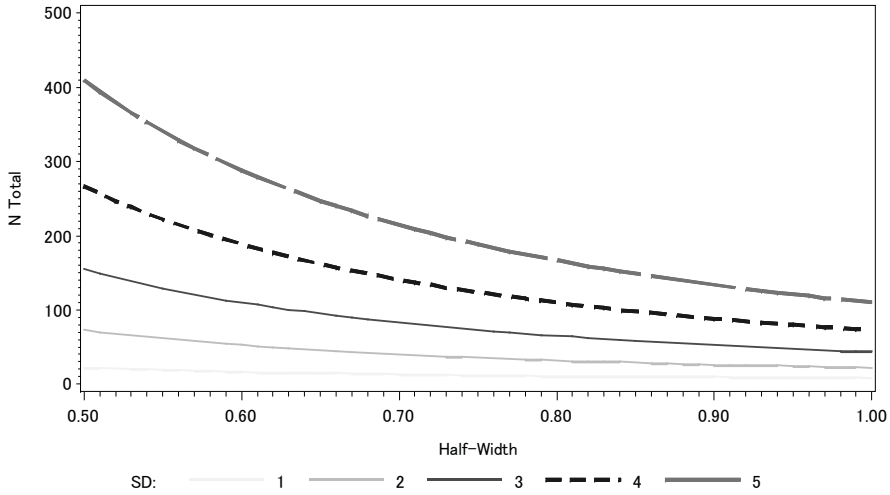


Figure 7. Relationship between half-width of 95% confidence interval for mean and sample size

Example 1:

Type of study: Non-intervention study

Objective of the study: determination proportion of patients with dyslipidemia who achieve target values of LDL-cholesterol after 12 month of therapy.

Sample size justification: Analysis of 3400 patients enables to determine the proportion of patients achieved the target values with precision (i.e. width of 95% Wald confidence interval) 3.36%. Taken account the worst scenario that only 50% of enrolled patients will be possible to include into analysis – the width of confidence interval would be 4.74%. (Results of the study published by Hradec et al.).

5. Factors influencing the power of statistical analysis

As was explained above the sample size is mainly affected by the following factors:

1. Level of significance
2. Power
3. Effect

The main factors which affect the power of the statistical test are the following:

Factors depending on type of data:

1. Variability of data (for continuous data)

2. Proportion in reference group (for binary data)
3. Incidence of the event (for censored data)
4. Type of data (categorical, censored, continuous)
5. Type of test (parametric vs. non-parametric)

Factors depending on design:

6. Ratio of number of subjects in groups to be compared
7. Parallel vs. cross-over design

5.1. Continuous data

Elements needed for sample size/power calculation if continuous parameter is planned to be compared, e.g. using by two sample t-test:

- Effect = difference in mean values of compared groups
- Variability of data = standard deviation
- Type of test (one sided, two sided, for equal/unequal SD)

Variability of data decreasing power of the test, implying more subjects is needed to include into analysis to achieve significant results. The relationship is not linear however depends on the effect size (Figure 8).

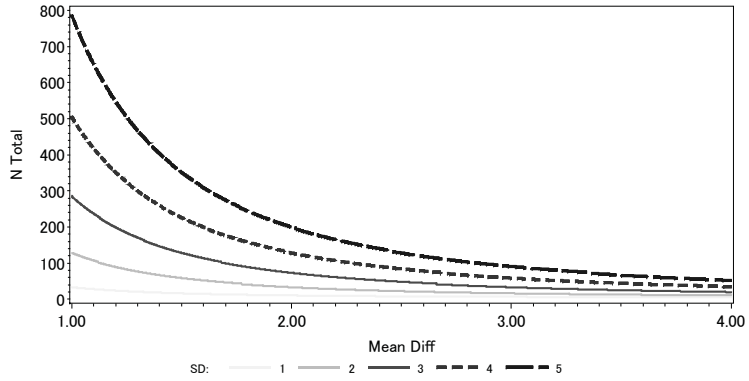


Figure 8. Relationship between effect and sample size for fixed power 80% of t-test and various SD

Example 2 (modeled study):

Type of study: Non-intervention study, non-inferiority study

Objective of the study: to compare levels of glucose in two groups of diabetic patients according baseline characteristics.

Sample size justification: The total number of 410 patients needs to be analyzed in order to achieve 90% power to demonstrate that group A is not inferior to group B in change of levels of glucose. According to clinical meaningful difference the non-inferiority margin was set to 1.5%. The sample size was established under assumption that difference between groups is 0.5% and standard deviation 3%. Taken account 10% drop-out the total number of patient planned to be enrolled into the study is 456 patients.

Selection of more specific subgroup of the study population would give us assumption of lower standard deviation (e.g. 2%). Given this assumption the total numbers of patients are 140 patients analyzed and 156 enrolled.

5.2. Binary data

Elements needed for sample size/power calculation if binary data are compared:

1. Effect = proportions in both groups
2. Type of test (one sided, two sided, chi-square, fisher, …)

While power of t-test depends on effect size (the difference of means) regardless the mean values in compared groups the power of chi-square test depends on both – the effects size (the difference of proportions) and proportion in reference groups. Towards to 50% in reference groups higher number of patients is needed to demonstrate the significant difference (Figure 9).

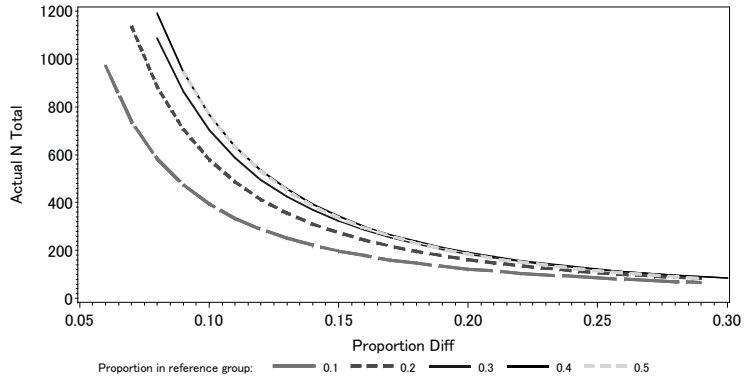


Figure 9. Relationship between effect and sample size for fixed power 80% of chi-square test and various proportions in reference group

Example 3 (modeled study):

Type of study: Superiority study

Objective of the study: to compare proportion of patients with improvement after treatment with study drug vs. placebo.

Sample size justification: The total number of 712 patients needs to be analyzed in order to achieve 80% power to demonstrate that proportion of improved patients treated with study drug is superior to those treated by placebo if the following assumptions are fulfilled: proportion of patients with improvement in study drug group 40% and in placebo group 30%.

In another study the proportions of patients with improvement are assumed to be higher 40% vs. 50% but the same treatment effect of 10% could be expected. The total number of patients to be enrolled is 776 patients.

5.3. Censored data

Elements needed for sample size/power calculation if censored data are compared using log-rank test:

1. Effect = reduction of incidence of the event
2. Incidence in reference group
3. Duration of the follow-up
4. Duration of accrual time (e.g. period of enrolment)
5. Type of test (one sided, two sided)

For power and sample size of the censored endpoint the number of events is the most important.

Alternatively,

1. Effect = median time of "survival" in both groups
2. Duration of the follow-up
3. Duration of accrual time (e.g. period of enrollment)
4. Type of test (one sided, two sided)

For censored data, number of events is the most critical feature in analysis of censored data. Therefore, incidence of the event in the study population and duration of follow-up change the power of log-rank test and sample size (Figure 10 and Figure 11, respectively). Effect of both incidence and duration of follow-up on the sample size is presented in Figure 12.

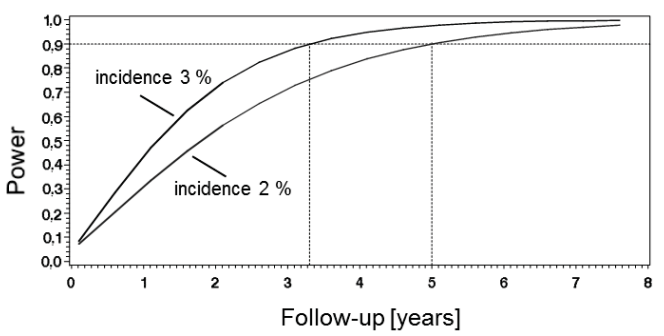


Figure 10. Relationship between power of log-rank test and duration of follow-up for various incidences of the event and fixed sample size. Source: Kadlecová 2009

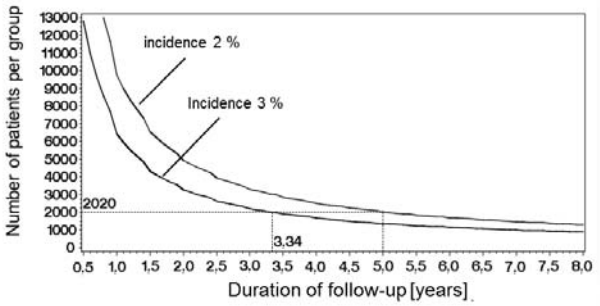


Figure 11. Relationship between sample size and power of log-rank test for various incidences of the event. Source: Kadlecová 2009

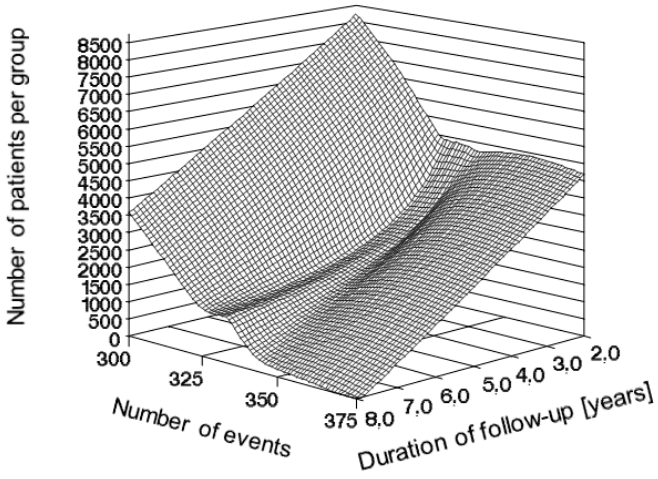


Figure 12. Relationship between sample size and incidence of event and duration of follow-up. Source: Kadlecová 2009

Example 3:

Type of study: ROADMAP: The Randomised Olmesartan And Diabetes MicroAlbuminuria Prevention study (Haller et al., 2006)

Objective of the study: to demonstrate decreasing incidence of mikroalbuminuria using by olmesartan medoxomil comparing to placebo

Given the assumption of 2% incidence of mikroalbuminuria , 30% reduction of the incidence by using the study drug the duration of the study was planned was planned to 5 years and 2020 patients per group was planned to be enrolled in order to achieve 90% power of log-rank test.

Results of interim analysis after 3 years shows that incidence of mikroalbuminuria is 3%. The decreasing of sample size was not possible as the patients had been already enrolled.

However, it was possible to shorten the follow-up. With given sample size 2020 patient per group and 3% incidence the 90% power of log-rank test is achieved after 3.34 (Figure 12.)

5.4. Type of data (categorical, censored, continuous)

Type of the parameter to be analyzed is very important factor which has effect on the statistical power. As the continuous data include the most information they are the most powerful. For comparison of binary parameters the most number of patients is needed. It is quite usual that sample size needed for binary and censored endpoint is only slightly different.

Example 4 (modeled study):

In order to demonstrate the effect of the type of primary endpoint to sample size let's suppose a model parallel study with active control which objective was to demonstrate effect of new treatment on decreasing of BMI after 12 month of therapy.

The following primary endpoints could be taken account:

a. *Change from baseline in BMI (i.e. continuous data)*
b. *Proportion of patients with decreasing by 4 kg/m^2 (i.e. binary data)*
c. *Time needed to achieve the first decreasing by 4 kg/m^2 (i.e. censored data)*

Using the real data from the study we can determine the sample size needed to achieve 80% power for the endpoints defined above.

a. *Change from baseline in BMI (i.e. continuous data)*
   - *Mean (±SD) values in test and reference group were 5.13 (±2.69) and 3.67 (±1.73), respectively.*
   - *The sample size needed for demonstration of significant difference between treatment groups in the change from baseline in BMI is 77 subjects in total.*
   - *Presented sample size was calculated for t-test for unequal SD in compared groups.*
b. *Proportion of patients with decreasing by 4 kg/m^2 (i.e. binary data)*
   - *The proportions of subjects with decreasing by 4 kg/m^2 in test and reference group were 69.39 % and 42.72 %, respectively.*
   - *The corresponding sample size is 108 subjects in total.*
   - *Presented sample size was calculated for chi-square test.*
c. *Time needed to achieve the first decreasing by 4 kg/m^2 (i.e. censored data)*
   - *The median time needed to achieve decreasing by 4 kg/m2 in test and reference group were 166 and 331 days, respectively.*
   - *The corresponding sample size is 102 subjects in total.*
   - *Presented sample size was calculated for log-rank test.*

5.5. Ratio of subjects in groups

It is not necessary to have number of subjects in both groups the same. Ratio 1:2 or 1:3 can be reasonable (e.g. for ethical reasons) but it decreases the power of test. Thus, more subjects

will be needed in total but in e.g. placebo group will be only a half of subjects. The increasing of total number of subjects in the study is graphically presented in Figure 13.

In epidemiological research it is usual that the groups are not balanced. If data with very different counts of subject in compared groups are analyzed the statistician should be aware that the power and reliability of the results is decreased.

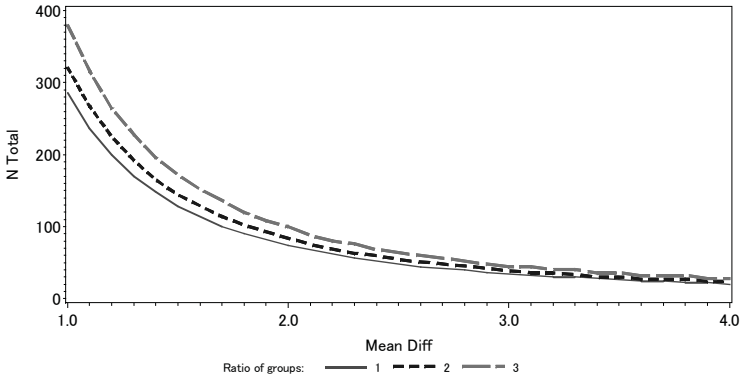


Figure 13. Relationship between effect and sample size for fixed power 80% of t-test and various ratios of groups

5.6. Parallel vs. cross-over design

Paired data obtained by e.g. cross-over study are more powerful to demonstrate the difference than comparing of independent samples. The sample size is not lower only about half. In case of t-test, the paired design can decrease the sample size approximately four-times (Figure 14).

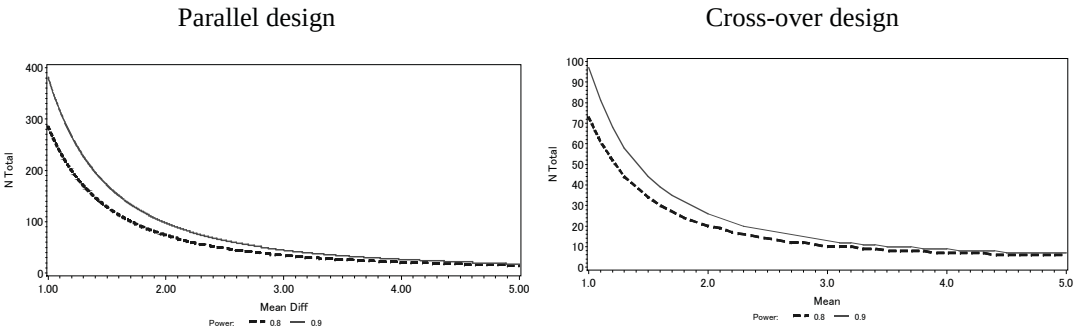


Figure 14. Relationship between effect and sample size for fixed power of t-test and parallel vs. cross-over design

6. Cross-validation and simulation in power analysis and sample size estimation

Epidemiological data could be used for sample size estimation of clinical trials. For sample size calculation we need assumptions for about the effect and about characteristics of the study population. We can use epidemiological data or data from previous studies of the population of interest, simulate supposed effect and calculate power directly from data.

Power of statistical test is probability of rejection H_0 if it is really false. Repeating the test e.g. 1000times we would be able to determine time the exact power for given sample size. Further, the power calculated by the statistical software could be cross-validated by using the epidemiological data.

Repeating of power calculation for different sample sizes we can obtain relationship between power and sample size for the statistical methods which could not be directly calculated using by statistical software. Useful e.g. for advanced statistical methods or non-parametric test.

7. References

European Medicines Agency. 2000. Points to consider on switching between superiority and non-inferiority. CPMP/EWP/482/99.

Suvarna V. 2010. Phase IV of Drug Development. Perspectives in Clinical Research 1: 57

Hradec J, Bultas J, Kmínek A, Hlaváč V, Tylová R, Kadlecová P. 2011. How statins are used in the Czech Republic? Results of observation study STEP [in Czech]. Cor et vasa 10: 527-534.

Kadlecová P. Analysis of the effect of study design on the sample size calculation. Brno, 2009. Diploma thesis. Masaryk University. Faculty of Science. Supervisor Adam Svobodník.

Haller H, Viberti G, Mimran A, Remuzzi G, Rabelink A, Ritz E, Rumpl L, Ruilope L, Katayama S, Ito S, Izzo J, Januszewicz A. 2006. Preventing microalbuminuria in patients with diabetes: rationale and design of the Randomised Olmesartan and Diabetes Microalbuminuria Prevention (ROADMAP) study. Journal of Hypertension 24: 403-408.

Basic aspects of clinical data management

Jaroslav Koča

ADDS s.r.o., Brno; e-mail: jkoca@adds.com

Abstract

The purpose of this work is to explain the basis of clinical data management to the students of computational biology and provide them elementary knowledge about this activity as it might be one of their possible areas of interest and employment. We will speak about the definition of clinical data management, the prerequisites of successful data management and its particular components.

Key words

Clinical trial, clinical data management, protocol, case report form

1. What is clinical data management

1.1. Definition

Clinical data management is a process to capture and transform the raw output from clinical trials into a usable form for statistical analysis and reporting.

It is very important to realize that good credibility and correctness of clinical study result strictly depends mainly on data. Thus, data management containing data capture, data processing, data validation and data storage strongly participates on the clinical trial results.

1.2. Objectives

The objectives of good clinical data management should be:

- To collect all relevant data
- To clean up all discrepancies and conserve the original information
- To assure the quality of collected data
- To provide accurate data in proper format for statistical analyses
- To store data for eventual review or further evaluation

2. Crucial information before starting with data management

2.1. Position of data management in terms of whole clinical study process

We can generally say that the sooner the data manager is involved in the project the better it is for the project. The ideal situation is when data manager can participate on the creation of study protocol. This is the moment when all the activities planned in the study are prepared, discussed and consequently fixed.

The first possible action performed under the responsibility of data manager is creation of case report form. From this step, through data transfer, data entry, data cleaning and validation to final export and archiving we talk about data management process.

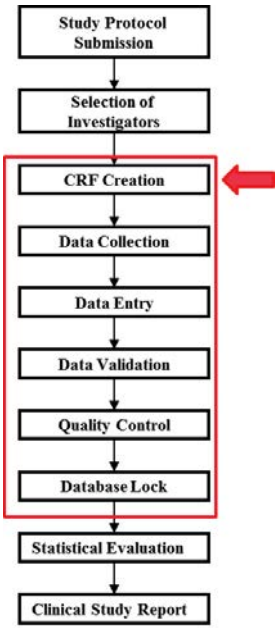


Figure 1. Position of data management in context of whole clinical study process

2.2. Data flow

Before starting with data management it is important to think about number and structure of data planned to be captured and processed. Besides crucial data displayed in case report forms, we can have specific sheets with laboratory data coming from local laboratories, set of laboratory data coming from central laboratory – usually transferred electronically and several inputs provided directly by patients like quality of life questionnaires or patient diaries. It is very important to know which sort of data are required and for what purposes they are captured. Such information is very helpful for successful setup of a clinical database and related coherence checks.

2.3. Important documents

There are two major documents for data management process. The first of them is study protocol containing all clinically relevant information, basic aspects of planned statistical analyses, the list of data to be collected, etc. Study protocol shall be followed during whole study process as the clue in case of some doubts.

The second important document is case report form (CRF). CRF is basic tool for complete data management process as a source for creation of clinical database and consecutive steps.

Data manager should take care to be in accordance with both study protocol and CRF during whole study process.

2.4. Standard operating procedures

Standard Operating Procedures (SOPs) are set of documents specific for each company and contain written instructions ensuring integrity of all performed activities. It is set of general rules describing how to perform particular tasks in each single activity. These instructions are in line with good clinical practice and contain link to all controlled documents used as templates for each particular project during the data management process.

2.5. Triangle principle

During the process of data management it is important to keep in mind continuous approach, so called triangle principle. This principle presents continuity of process in three different styles but in tight relation. At the first angle of triangle we can imagine activity (creation of database, programming of check, etc.), at the second angle we can imagine document describing this activity (e.g. document called Validation Plan describing all the checks to be applied on entered data), at the third angle we can imagine SOP giving rules how to perform the activity. It is important to realize the link between each two angles – connection like in the triangle:

- SOP – Template: Each template has link to at least one SOP.
- SOP – Activity: Each activity has background in at least one SOP.
- Activity – Template: Each activity is documented, can be reviewed, validated and eventually reconstructed.

3. Clinical data management process

The clinical data management process consists of many activities as shown in Figures 2 and 3 below. We can divide the global process into four main areas according to the status of project.

3.1. Setup activities

3.1.1. Project setup

Before doing any action we must be sure that we exactly know what to do and how to do it. As mentioned before we need knowledge about data flow and we should dispose of the two crucial documents, study protocol and CRF. Then we are able to dialogue with different people involved in the study and manage specific steps for successful setup of clinical database.

3.1.2. Database setup

Another and more concrete step is setup of clinical database. Particular details of this process may differ according to the computational system used, type of data entry preferred for current project, phase of clinical trial or client's requirements. However, the general objective is still the same: building a robust database corresponding CRF structure, matching to parameters defined in the protocol and containing required functionalities. Following necessities have to be assured:

- Database content. It must be assured that all relevant data are collected.
- Database structure. Particular variables must be organized in a way enabling easy and transparent data entry and comfortable data processing and analyses.

- Database validation. Database must be reviewed, tested and validated for both content and structure before provided for use in practice.

3.2. Data entry

Data entry part is relatively easier one from the organizational viewpoint. If project and database are prepared well, there is no need to modify anything and data are entered into database according predefined rules. There are two basic options how to enter data into database.

- Double data entry. This way of data entry is used for the most of clinical trials where paper CRF is used. The advantage of this option is that all the data are entered into database twice, by two independent people, which strongly minimize the possibility of mistake.
- Simple data entry. This option is used for less important trials with lower budget (e.g. post marketing or non-intervention trials) or in trials where electronic data capture (EDC) is used. EDC represents approach where investigators enter data directly into the clinical database usually via internet browser interface.

3.3. Data cleaning

Data cleaning represents very complex set of activities whose main objective is to review data entered into the database from many viewpoints to ensure their correctness and validity. It is a process where eventual inconsistencies found are solved with investigators via queries, signed controlled forms to document all modifications in data. Some of the data cleaning steps are:

- Control of format checking whether entered value corresponds to the predefined format of each variable (defined length of field, minimal/maximal permitted value, coding options – e.g.: 1 = male, 2 = female).
- Control of coherence checking coherence between two and more variables (required distance between particular visits, matching between inclusion criteria and related parameters, etc.)
- Medical coding is the process to standardize mentioned concomitant medication, medical history or adverse event terms. Text fields containing particular terms are coded according standard dictionaries and then much more easily processed and analyzed.
- Medical review is the process to review data from medical point of view. This step should be performed by medical expert to find inconsistencies among similar information mentioned on different part of CRF.
- Translation could be used in multi-country trials where text field terms are often displayed in local language. Unification into one language is usually a standard process.
- SAE reconciliation is an important process for most of trials. The point of this process is to reconcile two databases containing information about serious adverse events occurred during the study duration. Each serious adverse event (SAE) has to be captured in clinical database (together with other study data) and also in safety database focused on reporting of SAEs. Not only the number of SAEs but also their character has to be the same.

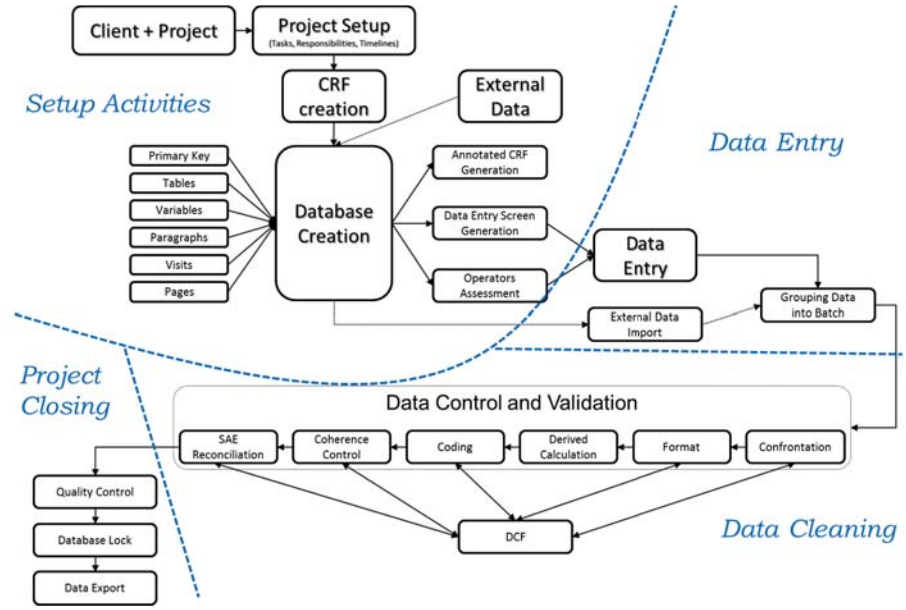


Figure 2. Data Management Process using paper case report forms

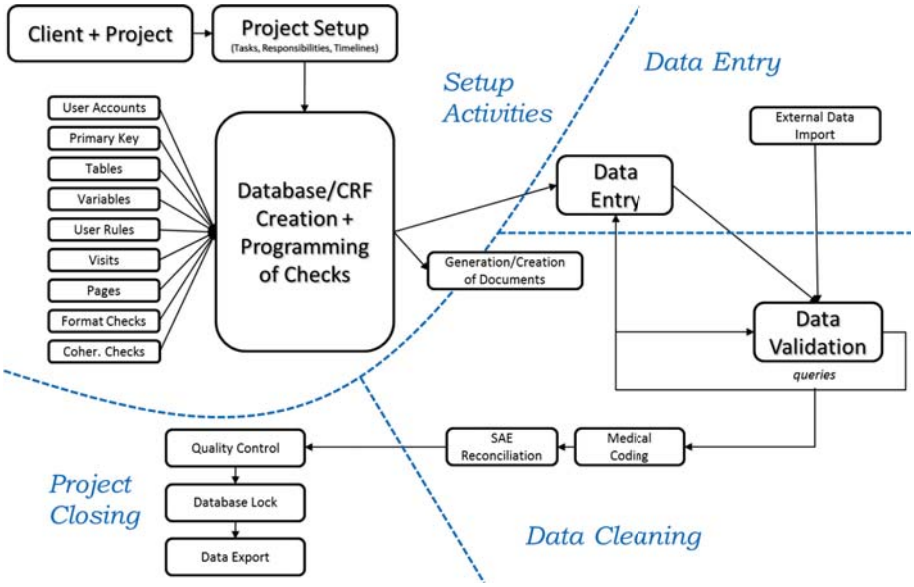


Figure 3. Data Management Process using electronic data capture system

3.4. Project closing

Study closing activities from the data management viewpoint are starting once all the inconsistencies are found and all the queries sent to investigators are resolved. The boundary between data cleaning and study closing is on data review meeting. Data review meeting is very specific and very important meeting organized before locking the database to discuss all

aspects of data, to ensure that all the inconsistencies are solved and to define and fix study populations for statistical analyses. After this meeting once the last queries are resolved, the clinical database can be locked. From this moment no modification in database can be done unless specific requirements are asked by the client. Eventual database unlock has to be documented properly giving clear reasons for unlock. Anyway, locking of clinical database is the last step of data management working with data. The remaining activities consist only of documentation and archiving processes.

4. Conclusion

Clinical data management is very complex process containing a lot of activities. As all the activities need to be carefully planned and scrupulously met, strong project management skills are required for successful data management of important clinical trials. It is necessary to stay kept in touch with all the key people involved in the study and continuously keep in mind project status and timelines. Only once the database is locked, data manager can slow down and start preparing materials for archiving.

Estimating number of cancer patients potentially treated with anti-tumour therapy

Tomáš Pavlík, Ondřej Májek, Jan Mužík, Ladislav Dušek

Institute of Biostatistics and Analyses, Masaryk University, Brno;
e-mail: pavlik@iba.muni.cz

Abstract

The objective of this paper is to present a model for estimation of time period prevalence of cancer patients requiring anti-tumour therapy. Besides incidence estimate, the model also provides the estimates of patients with terminal and non-terminal form of cancer. Moreover, the whole estimation process should be accessible from population-based cancer registry data. The proposed method has been designed with respect to the extent of cancer because for many types of cancer the clinical stage is by means of patients' life-expectation and anticipated financial costs of the treatment even more influencing than age at diagnosis. To document its applicability, the model was applied on colorectal cancer data from the Czech National Cancer Registry to model the number of potentially treated patients with colorectal cancer in the Czech Republic in 2015.

Key words

Statistics, modelling, cancer, survival analysis.

1. Introduction

Modern anti-tumour therapy introduces significant improvement in survival of cancer patients, therefore, leading to increasing cancer incidence and prevalence rates (Ferlay et al., 2010). Such a progress introduces the essential need for monitoring and prospective planning of number of patients eligible for targeted therapy, as necessary financial resources need to be allocated. Estimation of cancer incidence and prevalence can only be seen as the first step in the process focused on the potentially treated patients as the prevalence estimates need to be further adequately adjusted for patients untreated from whatever reason (cure for cancer, treatment contraindication, very high age, patient's refusal to treatment, advanced stage of disease). In any case, the cancer prevalence estimation is not an easy task because it cannot be estimated directly from the population-based data due to time limited registration of the cancer cases, and has to be modelled. Moreover, the model should be designed with respect to the extent of cancer because the clinical stage is by means of patients' life-expectation and anticipated financial budget impact of the treatment even more influencing than age at diagnosis (Clerc et al., 2008). The existing models use either only population data (Mariotto et al., 2006) or a combination of population data and clinical records (Gatta et al., 2004; Chauvenet et al., 2009). In the former case, the model does not employ a concept of cancer recurrence, whereas in the latter case, the concept of cancer recurrence is considered and the particular rates are estimated from the hospital records.

The objective of this paper is to present a model for estimation of time period prevalence of cancer patients requiring anti-tumour therapy. To document its applicability, the model was applied on colorectal cancer (CRC) data from the Czech National Cancer Registry (CNCR)

to model the number of potentially treated patients with colorectal cancer in the Czech Republic in 2015.

2. Patients and Methods

2.1. Patients

The Czech Republic makes use of high-quality population-based data on cancer epidemiology:

- CNCR (data administrator and provider: Institute of Health Information and Statistics of the Czech Republic, IHIS) covers the whole population of the Czech Republic (10,230,000 inhabitants according to the 2001 census) since 1976. Reference dataset defined for the period 1995-2008 (which is more relevant for recent development) involves records on more than 500,000 patients.
- Demographic data on the Czech population and the Death Records Database (data administrator and provider: Czech Statistical Office, CZSO) constitute an indispensable background information for the predictive assessment of epidemiological data. This data was used for the adjustment of age-standardized predictions of incidence rates.

Regarding the analysis of CRC data, a total of 179,286 incident CRC cases (12% of the CNCR records) were registered in the CNCR in the period 1982-2008. Data on cases diagnosed in 1977-1981 were excluded due to the lack of a classification system for clinical stages. Moreover, only clinically relevant cancer records entered the modelling procedures. The epidemiological records on patients diagnosed by death certificate only or at autopsy were excluded from the analysis. Finally, 160,017 incident cases were considered for the analysis. Four age categories were considered in the modelling: 0-49 years, 50-64 years, 65-79 years and 80+ years; as well as three categories for the disease extent: clinical stages I and II, representing localised CRC, clinical stage III, representing regionally advanced disease, and clinical stage IV, representing metastatic disease. Colorectal cancers diagnosed in stage I or II were merged prior to analyses due to changes in the TNM classification system (Hermanek and Sobin, 1992). Moreover, cases with missing information on stage (denoted here as X) were also considered for the model, as they represent an indispensable mass of patients that needs to be accounted for in the health care system.

2.2. The model

The model was described in detail elsewhere (Pavlík et al., 2012). It comes from the model of period prevalence defined as the proportion of patients with present or past diagnosis of cancer alive in a population in a certain year. The modelling process has two steps. In the first step, overall number of living cancer patients irrespective of the anti-tumour therapy applied is identified. The prediction combines the number of newly diagnosed patients and the number of patients who were diagnosed previously and lived at the year of interest. In the second step, number of patients probably treated in a given year due to a primary disease or due to a recurrence of the primary disease is estimated. As mentioned previously, the model is derived in a stage-specific manner as this stratification is necessary in a case of financial planning since the treatment costs and other resources needed are highly associated with the cancer stage. Moreover, several scenarios can be adopted to cover the plausible development of the incidence and survival rates, and the probability of an anti-tumour therapy initiation.

Considering the extent of cancer, s, as the main stratification factor for the estimation of cancer prevalence, we will categorise it into three groups according to clinical stages defined

by the TNM classification system: s = I + II for clinical stages I and II (representing localised disease); s = III for clinical stage III (representing regionally advanced disease); and s = IV for clinical stage IV (representing metastasized disease). The stage-specific prevalence, $P_s(y)$, can be then expressed as follows:

$$P_s(y) = \sum_{a=1}^{m} P_{s,a}(y) = \sum_{a=1}^{m}\sum_{i=0}^{n} I_s(y-i,a)S_s(i,a), \tag{1}$$

where a is a categorical age cohort variable of m categories and $P_{s,a}(y)$ denotes the prevalence of patients ever diagnosed at ath age category and stage category s alive in calendar year y. The $P_{s,a}(y)$ can be further formulated as the convolution of incidence and survival functions: $I_s(y - i, a)$ and $S_s(i, a)$ are the age and stage-specific incidence and survival functions, respectively, and n is the number of annual incidence figures available for the computation.

Equation (1) can be easily split into two terms (assuming newly diagnosed patients being prevalent in the year of interest and thus having $S_s(0, a) = 1$) as follows:

$$P_{s,a}(y) = \sum_{i=0}^{n} I_s(y-i,a)S_s(i,a) = I_s(y,a) + \sum_{i=1}^{n} I_s(y-i,a)S_s(i,a), \tag{2}$$

First term on the right-hand side of equation (2) represents the newly diagnosed patients whereas the second one represents patients diagnosed in the past and alive in the given year. Correcting the first term of (2) for the probability of being untreated with anti-tumour treatment due to poor health condition or other objective reasons (e.g. patient's refusal) and simultaneously correcting the second term of (2) in a way that only patients with the recurrence of the disease in a good health condition allowing the anti-tumour treatment are considered, the prevalence of patients receiving active anti-tumour therapy, denoted as $P^*_{s,a}(y)$, can be derived as follows:

$$P^*_{s,a}(y) = I_s(y,a)\delta_s(y,a) + \sum_{i=1}^{n} I_s(y-i,a)S_s(i,a)R_s(i,a)\delta_s(y,a), \tag{3}$$

where $\delta_s(y, a)$ is the stage- and age-specific probability of being treated with an anti-tumour treatment in the year of interest and $R_s(i, a)$ is a function that describes the risk of suffering from cancer recurrence after surviving i years from diagnosis.

The cancer recurrence function, $R_s(i)$, need to be further specified using the following consideration: each patient diagnosed in stage s can suffer in time from two forms of cancer recurrence, either non-terminal, actually not leading to death in the year y, denoted as $R_s^1(i)$, or terminal, leading to death in the year y, denoted as $R_s^2(i)$. The stratification further determines the patient's treatment course. In the former case, the patient is assumed to be treated in a similar way as at the time of primary diagnosis, i.e. the patient stays in the prevalence pool of the particular stage s. In the second case, the patient is assumed to be treated for generalized (or metastasized) disease, i.e. the patient moves from the prevalence of stage I+II or III to the prevalence of stage IV.

Splitting the $R_s(i)$ term in equation (3) and moving the patients suffering from terminal cancer recurrence to the prevalence of stage IV led to the following formulation of the stage-

specific prevalence of patients requiring active anti-tumour therapy (for simplicity, index *a* representing the age category is omitted):

$$P_s^*(y) = I_s(y)\delta_s(y) + \sum_{i=1}^{n} I_s(y-i)S_s(i)R_s^1(i)\delta_s(y); \quad s = \mathrm{I+II, III, X}, \tag{4}$$

$$\begin{aligned} P_{\mathrm{IV}}^*(y) &= I_{\mathrm{IV}}(y)\delta_{\mathrm{IV}}(y) + \sum_{i=1}^{n} I_{\mathrm{IV}}(y-i)S_{\mathrm{IV}}(i)\left(R_{\mathrm{IV}}^1(i) + R_{\mathrm{IV}}^2(i)\right)\delta_{\mathrm{IV}}(y) \\ &+ \sum_{s=\mathrm{I+II,III,X}} \sum_{i=1}^{n} I_s(y-i)S_s(i)R_s^2(i)\delta_{\mathrm{IV}}(y). \end{aligned} \tag{5}$$

2.3. Specifying colorectal cancer incidence, survival, and recurrence

In the proposed model, the age-drift Poisson regression model was applied for estimating CRC incidence employing two different link functions: the identity link function was used to model increasing incidence trends, whereas the logarithmic link was utilised to model decreasing trends (Dyba and Hakulinen, 2000). Two scenarios can be considered for the estimation of incidence rates. First, CRC incidence rates can be considered fixed at the values observed in 2008; second, the age, period and cohort model can be applied for the estimation of future counts (Bray and Moller, 2006).

The stage-specific survival rates were estimated using a method based on the moving window principle that employs the standard life-table method (Marubini and Valsecchi, 2004). In this procedure, the survival rates are estimated successively, using the cohort analysis of patients diagnosed in overlapping 5-year time intervals. To ensure validity, calculation of *x*-year survival rates is only performed on cohorts in whom the *x*-year survival rate can be reliably estimated, and were diagnosed as recently as possible (Dušek et al., 2009). Two scenarios can be adopted for survival estimates as well. In the first scenario, the survival rates can be assumed to improve from 2008 to 2015 in the same manner as observed in the CNCR data from the period of 2004-2008. In the second scenario, survival rates needed for calculating the 2009-2015 prevalence are fixed at the most recent values, namely the survival rates available in 2008.

The records of cancer recurrence rates are not directly available on the population level in the Czech Republic. For this reason, surrogate parameters were used to estimate the cancer recurrence rates. Regarding non-terminal cancer recurrence rates, $R_s^1(i)$, these were estimated using the information on the patient's health status and non-symptomatic anti-tumour therapy applied after the first year following diagnosis (first year after diagnosis is assumed to correspond to the initial treatment phase). However, as the $R_s^1(i)$ function refers only to non-terminal cancer recurrence, there was an additional condition needed and that was that the patients had to survive up to the end of the particular year of interest, i.e. the cancer recurrence had not to be terminal in a given year.

As for terminal cancer recurrence, the $R_s^2(i)$ function was estimated using the information on cancer as the main cause of death in the CNCR. The approach is based on the assumption that nobody can die from cancer, with cancer being the main reason of death, without passing through the phase of generalized disease. Therefore, the $R_s^2(i)$ function represents the excess mortality of the cancer and can be thus specified using either the relative survival function or the underlying excess hazard rate (Dickman et al., 2004).

As the last factor needed for the model, the proportion of patients treated with anti-tumour therapy reflecting the patients' health status were derived from the CNCR population data. Like in the case of CRC incidence and survival rates, two scenarios can be considered for the proportion of the treated CRC patients. First, this proportion can be regarded fixed and estimated in a stage-specific manner from the period 2004-2008. Second, the values observed from the CNCR can be extrapolated forward in time using a logistic regression model.

For the sake of completeness, the eight models considered in this paper that are defined by combination of two scenarios for the estimation of incidence rates (fixed and modelled, respectively), two scenarios for the estimation of survival rates (constant and improving, respectively), and two scenarios for the proportion of treated patients (fixed and modelled, respectively) are summarised in Table 1.All computations were performed using Stata 10.1 software.

Table 1. Description of the eight scenarios used to estimate the number of colorectal cancer patients treated with anti-tumour therapy in 2015 in the Czech Republic

Proportion of treated patients (for the year 2015)	Incidence rates (for the period 2009-2015)	Survival rates (for the period 2009–2015)	
		Survival rates are considered fixed at the most recent values, i.e. survival rates in 2008	Survival rates are assumed to improve in the same manner as observed in the period 2004–2008
Proportion is regarded fixed in time and estimated from the period 2004–2008	Incidence rates are considered fixed at the values observed in 2008	*Scenario 1*	*Scenario 2*
	Incidence rates are modelled in time using the age-drift Poisson regression model	*Scenario 3*	*Scenario 4*
Proportion observed in the period 2004–2008 is extrapolated forward in time using a logistic regression model	Incidence rates are considered fixed at the values observed in 2008	*Scenario 5*	*Scenario 6*
	Incidence rates are modelled in time using the age-drift Poisson regression model	*Scenario 7*	*Scenario 8*

3. Results

The results were described in detail elsewhere (Pavlík et al., 2012). Applying the selected scenarios, the 2015 CRC prevalence of patients primarily diagnosed in stage I or II is estimated as ranging between 338.8 and 389.8 per 100,000 people, while the prevalence of patients diagnosed in stage III is estimated as ranging between 114.1 and 150.2 per 100,000 people, and the prevalence of patients diagnosed in stage IV ranging between 50.7 and 58.1 per 100,000 people. The prevalence of CRC patients with missing information on stage in CNCR is estimated as ranging between 26.3 and 33.9 per 100,000 people. As expected, the biggest discrepancy between the scenarios can be seen for the merged stages I+II where the improvements in survival are manifested the most. In total, between 529.9 and 632.0 CRC patients per 100,000 people are estimated to be prevalent in 2015. The model thus predicts an increase in CRC prevalence from 13% to 30% in comparison with the situation in 2008. This increase underlines the seriousness of the CRC burden in the Czech Republic.

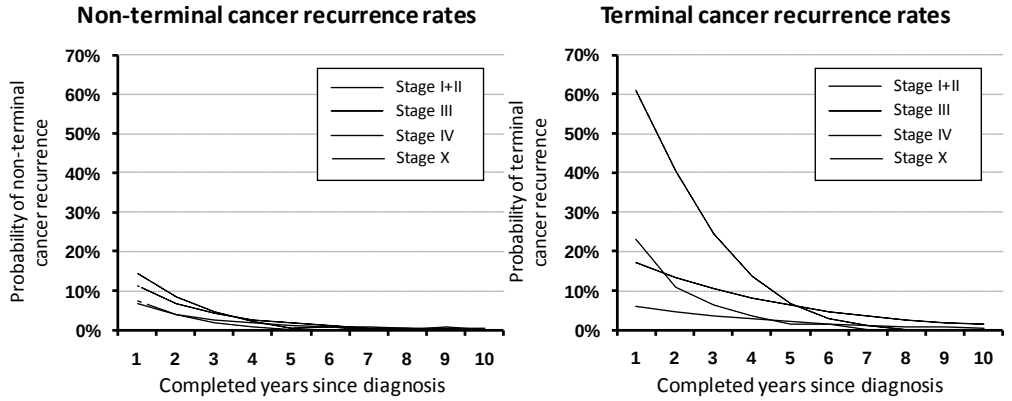


Figure 1. Stage-specific estimates of non-terminal and terminal recurrence rates in first ten years after primary diagnosis of colorectal cancer; the estimates correspond to the recent time period, 1995-2008

Figure 1 shows the estimated stage-specific rates of non-terminal and terminal cancer recurrence, respectively; in the ten years following the first completed year after diagnosis (first year after diagnosis is considered to correspond to primary therapy). The estimates corresponding to the most recent time period, 1995-2008, are shown. We can see the risk of non-terminal cancer recurrence gradually decreasing in the first three years and reaching the 3% level in all stages afterwards. On the contrary, the pattern of terminal recurrence rates varies with clinical stage up to five years after diagnosis; in stages I + II, the recurrence rates are consistently below 7%, conveying a good perspective of patients diagnosed with less advanced disease. In stage III, the terminal recurrence rate shows a stable but very slow decrease in time. The terminal recurrence rate for stage IV reveals a very high risk of dying from CRC exceeding even 60% after the first year following diagnosis. The risk reaches the level comparable to other stages after 5-6 years following the diagnosis. The terminal recurrence rate of the patients with missing information on stage is located in the middle of the other stage-specific profiles (Figure 1). It documents the fact that patients with missing information on stage represent a mixture of patients of all stages.

The numbers of patients requiring active anti-tumour therapy for the CRC in the Czech Republic in 2015 estimated according to eight considered scenarios are given in Table 2. For each scenario, the first three columns represent the individual components of the proposed model: the estimated number of newly diagnosed and treated patients, the estimated number of patients treated for non-terminal cancer recurrence, and the estimated number of patients treated for terminal cancer recurrence, respectively. Then, the sums with respect to the stage at the diagnosis are shown (column 4).

In total, from 10,074 to 11,440 CRC patients are predicted for anti-tumour therapy administration in the Czech Republic in 2015 according to the eight scenarios considered for incidence and survival rates and the probability of anti-tumour therapy administration. When regarding the stage at diagnosis as the primary stratification factor, 4,595 to 4,828 patients (41-47% of all CRC patients) primarily diagnosed in stage I or II; 2,679 to 3,613 patients (27-32%) primarily diagnosed in stage III; and 2,366 to 2,969 patients (23-27%) primarily diagnosed in stage IV are estimated to be treated in 2015, respectively. Regarding patients with missing information on stage, 134 to 335 of them (1-3%) are predicted for anti-tumour therapy in 2015.

4. Discussion

Modelling the prevalence of the CRC patients requiring active anti-tumour therapy is an important issue; especially in countries like the Czech Republic which ranks among countries with the highest cancer load worldwide (Ferlay et al., 2010). The stage-specific modelling is complicated and requires a comprehensive approach, since the stage at the time of diagnosis need not necessarily coincide with the disease extent several years afterwards. The disease extent should be taken into account in the modelling process at all time points because the clinical stage is, in regards to patient life-expectation and anticipated financial costs, even more important than age at diagnosis (Clerc et al., 2008). That is why we attempt to propose a comprehensive statistical method here that may provide such estimates in a stage-specific manner utilizing solely the population-based cancer registry data.

In accordance with other epidemiological studies, for example Gail et al. (1999), four extreme scenarios regarding progress in incidence and survival rates were implemented to model the CRC prevalence in this study. The incidence rates were either assumed fixed at the 2008 level or modelled using the age, period, and cohort model. As for the survival rates, they were either assumed to improve from 2008 to 2015 at the same rate as observed in the period of 2004-2008 or fixed at the most recent values, i.e. the survival rates available in 2008.

The estimated one-year prevalence rates are not directly comparable with the international data coming from comparative studies such as Engholm et al. (2010), since these studies focus primarily on the point prevalence. However, at least a crude comparison shows that the prevalence of CRC in the Czech Republic gradually reaches the situation in the Western and Northern European countries. A very high incidence rate and the already mentioned successively improving survival rates can be regarded as the two main drivers.

Table 2. Stage-specific estimates of prevalence of patients requiring active anti-tumour therapy for colorectal cancer in the Czech Republic in 2015 according to the eight scenarios – part 1.

Scenario 1: Constant proportion of treated patients; Constant incidence rate; Constant survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,650	565	479	4,694
Stage III	1,783	355	541	2,679
Stage IV	1,419	181	766	2,366
Missing	220	16	99	335
All cases	7,072	1,117	1,885	10,074

Scenario 3: Constant proportion of treated patients; Modelled incidence rate; Constant survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,581	547	467	4,595
Stage III	2,223	422	632	3,277
Stage IV	1,428	177	761	2,366
Missing	131	10	59	200
All cases	7,363	1,156	1,919	10,438

Scenario 5: Modelled proportion of treated patients; Constant incidence rate; Constant survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,613	560	562	4,735
Stage III	1,831	362	628	2,821
Stage IV	1,675	206	890	2,771
Missing	122	9	121	252
All cases	7,241	1,137	2,201	10,579

Scenario 7: Modelled proportion of treated patients; Modelled incidence rate; Constant survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,542	544	549	4,635
Stage III	2,285	429	734	3,448
Stage IV	1,697	203	887	2,787
Missing	72	5	71	148
All cases	7,596	1,181	2,241	11,018

Table 2. Stage-specific estimates of prevalence of patients requiring active anti-tumour therapy for colorectal cancer in the Czech Republic in 2015 according to the eight scenarios – part 2.

Scenario 2: Constant proportion of treated patients; Constant incidence rate; Improving survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,650	607	524	4,781
Stage III	1,783	407	625	2,815
Stage IV	1,419	212	898	2,529
Missing	220	13	76	309
All cases	7,072	1,239	2,123	10,434

Scenario 4: Constant proportion of treated patients; Modelled incidence rate; Improving survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,581	589	511	4,681
Stage III	2,223	475	725	3,423
Stage IV	1,428	207	892	2,527
Missing	131	9	47	187
All cases	7,363	1,280	2,175	10,818

Scenario 6: Modelled proportion of treated patients; Constant incidence rate; Improving survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,613	602	613	4,828
Stage III	1,831	415	727	2,973
Stage IV	1,675	241	1,038	2,954
Missing	122	7	92	221
All cases	7,241	1,265	2,470	10,976

Scenario 8: Modelled proportion of treated patients; Modelled incidence rate; Improving survival rates				
Stage at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total
Stage I+II	3,542	583	599	4,724
Stage III	2,285	484	844	3,613
Stage IV	1,697	236	1,036	2,969
Missing	72	5	57	134
All cases	7,596	1,308	2,536	11,440

Two principal types of estimates for the cancer recurrence rates are widely used, either estimates based on clinical or hospital data (Liang et al., 2007) or estimates coming from population-based databases (Manfredi et al., 2006). We feel that the estimates coming from the population-based databases may be more appropriate in this type of modelling, as the estimates calculated from hospital data can lead to biased results due to non-representativeness of the underlying set of patients. On the other hand, the precise information on time of cancer recurrence is barely available in the population-based cancer registries. In our model, the rationale behind the estimation of functions representing the non-terminal and terminal cancer recurrence rates, respectively, was to use surrogate parameters. The terminal form of cancer recurrence was estimated from the information on cancer as the main cause of death, whereas the non-terminal form was identified from the information on patient's vital status and anti-tumour therapy applied during the follow-up period. Of course, the need for the surrogate information introduces high requirements on the data quality of the population-based registry.

Considering the most recent changes in CRC epidemiology and care in the Czech Republic, we feel that the most likely scenario for the year 2015 is the one with stabilised incidence rates, improving survival rates, and an increasing proportion of treated patients (see Table 2, scenario 6). The stabilised incidence rates can be expected due to the increase in attendance at the national organised screening program that has been observed during very recent years in the Czech Republic (Májek et al., 2010). In addition, both the improvement in survival rates and the increasing proportion of treated patients can be attributed to the establishment of a network of highly specialised Cancer Centres that took place in the Czech Republic in 2006 (Fínek et al., 2009), and the introduction of molecular targeted therapy in recent years.

5. Conclusion

A model for the estimation of the number of cancer patients requiring active anti-tumour therapy in a stage-specific manner utilizing solely the population-based cancer registry data was proposed in this contribution. In total, eight scenarios concerning progress in incidence rates, survival rates, and the probability of an anti-tumour therapy administration were considered for the estimation of the number of potentially treated CRC patients. Based on the scenarios, the model predicted an increase in CRC prevalence ranging from 13% to 30% in comparison with the situation in 2008. The model also predicted that the number of colorectal cancer patients requiring active anti-tumour therapy in the Czech Republic in 2015 ranges from 10,074 to 11,440. Moreover, 3,485 to 4,469 patients will be potentially treated for the metastatic disease, which accounts for more than one third of all CRC patients.

6. References

Bray F, Møller B. 2006. Predicting the future burden of cancer. Nature Reviews in Cancer 6**:** 63-74.

Chauvenet M, Lepage C, Jooste V, Cottet V, Faivre J, Bouvier AM. 2009. Prevalence of patients with colorectal cancer requiring follow-up or active treatment. European Journal of Cancer, 45: 1460-1465.

Clerc L, Jooste V, Lejeune C, Schmitt B, Arveux P, Quantin C, Faivre J, Bouvier AM. 2008. Cost of care of colorectal cancers according to health care patterns and stage at diagnosis in France. European Journal of Health Economics 9: 361-367.

Dickman PW, Sloggett A, Hills M, Hakulinen T. 2004. Regression models for relative survival. Statistics in Medicine 23: 51-64.

Dušek L, Pavlík T, Májek O, Koptíková J, Gelnarová E, Mužík J, Vyzula R, Fínek J. 2010. Information System for Predictive Evaluation of Cancer Epidemiology and the Number of Cancer Patients in the Czech Republic. In: Dušek L, Ed. Czech Cancer Care in Numbers 2008-2009. Praha: Grada Publishing 2010: 255-273.

Dyba T, Hakulinen T. 2000. Comparison of different approaches to incidence prediction based on simple interpolation techniques. Statistics in Medicine 19: 1741-1752.

Engholm G, Ferlay J, Christensen N, Bray F, Gjerstorff ML, Klint A, Køtlum JE, Olafsdóttir E, Pukkala E, Storm HH. 2010. NORDCAN-a Nordic tool for cancer information, planning, quality control and research. Acta Oncologica 49: 725-736.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. 2010. GLOBOCAN 2008, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 Lyon: International Agency for Research on Cancer. Available from: http://globocan.iarc.fr.

Fínek J, Vyzula R, Petera J, Indrák K, Štěrba J, Vorlíček J, Dušek L. 2010. Network of Cancer Care in the Czech Republic. In Czech Cancer Care in Numbers 2008-2009. In: Dušek L, Ed. Czech Cancer Care in Numbers 2008-2009. Praha: Grada Publishing 2010: 24-32.

Gail MH, Kesser L, Midthune D, Scoppa S. 1999. Two approaches for estimation disease prevalence from Population-based registries of incidence and total mortality. Biometrics 55: 1137-1144.

Gatta G, Capocaccia R, Berrino F, Ruzza MR, Contiero P. 2004. Colon cancer prevalence and estimation of differing care needs of colon cancer patients. Annals of Oncology 15: 1136-1142.

Hermanek P: Sobin LH: International Union Against Cancer (UICC). 1992. TNM Classification of Malignant Tumors. 4 edition. Berlin: Springer 1992.

Liang JT, Huang KC, Lai HS, Lee PH, Jeng YM. 2007. Oncologic results of laparoscopic versus conventional open surgery for stage II or III leftsided colon cancers: a randomized controlled trial. Annals of Surgical Oncology 14: 109-117.

Májek O, Daneš J, Zavoral M, Dvořák V, Suchánek S, Seifert B, et al. 2010. Czech National Cancer Screening Programmes. Klinická Onkologie 23: 343-353.

Manfredi S, Bouvier AM, Lepage C, Hatem C, Dancourt V, Faivre J. 2006. Incidence and patterns of recurrence after resection for cure of colonic cancer in a well defined population. British Journal of Surgery 93: 1115-1122.

Mariotto AB, Yabroff KR, Feuer EJ, De Angelis R, Brown M. 2006. Projecting the number of patients with colorectal carcinoma by phases of care in the US: 2000-2020. Cancer Causes Control, 17: 1215-1226.

Marubini E, Valsecchi MG. 2004. Analysing Survival Data from Clinical Trials and Observational Studies. New York: John Wiley & Sons 2004.

Pavlík T, Májek O, Mužík J, Koptíková J, Slavíček L, Fínek J, Feltl D, Vyzula R, Dušek L. 2012. Estimating the number of colorectal cancer patients treated with anti-tumour therapy in 2015: the analysis of the Czech National Cancer Registry. BMC Public Health 12: 117.

Stochastic Modelling
in Epidemiology

Computational Biology Students' Abstracts

Estimation of relative survival of patients after PCI

Klára Benešová

Faculty of Science, Masaryk University, Brno; e-mail: 374222@mail.muni.cz

Abstract

Percutaneous coronary intervention (PCI) is helpful in treatment and prevention of ischaemic heart disease (IHD). We compared survival of patients after PCI with survival of the general Czech population to find out whether PCI patients decease less or more than are expected. Relative survival, the ratio of the observed and the expected survival, was used to estimate disease-specific survival. An analysis of the National Register of Cardiovascular Interventions (NRKI) showed absolute (observed) survival of 78.4 % patients (81.3 % in patients with survival > 30 days) and relative survival of 93.8 % patients (96.6 % in patients with survival > 30 days) at 5 years of follow-up. Relative survival was higher in men and patients ≥ 75 years.

Key words

Population-based survival analysis, PCI, NRKI, relative survival

1. Introduction

In 2010, ischaemic heart disease (IHD) caused 25 178 deaths (23.6 % of all deaths that year) in the Czech Republic (Eurostat, 2013). Percutaneous coronary intervention (PCI) is one of the ways how to treat IHD. PCI is a non-surgical procedure used to widen narrowed coronary arteries. A deflated balloon on a catheter is placed into the narrowed artery. After that, the balloon is inflated to open the artery and a stent is inserted at the site of blockage to keep the artery permanently open.

We focused on probability estimations of relative survival of patients after PCI to gain the information whether PCI gives an advantage in a future patient's survival. Relative survival is the ratio of the observed and the expected survival rates which gives an estimate of survival due to the disease of interest without the need of information on individual cause of death. To obtain the relative survival we compared the survival rate in PCI patients with that in the total Czech population, adjusted for sex, age and calendar time.

2. Methods

2.1 Patients

The National Register of Cardiovascular Interventions (NRKI) is the analysis-based data source which contains records of cardiovascular interventions performed in the Czech Republic from January 2005 to the end of September 2011. During this time period, 97,844 patients underwent PCI. From 120,419 records, we excluded repeated interventions and enrolled 86,386 patients in which their first PCI was performed in the considered time frame. This study group consisted of 58,881 men (68.2 %; mean age 63 years) and 27,505 women (32.8 %; mean age 70 years). Patients were divided into three age groups < 60, 60-74 and ≥ 75 (Table 1). For the purpose of the long-term survival analysis, we considered only 83,262 patients who were alive after 30 days from the intervention.

Table 1. Clinical characteristics of 86,386 patients

Characteristics	No. or mean ± SD	% or range
Age	65.6 ± 11.4	22-100
Sex		
Male	58,881	68.2
Female	27,505	31.8
Age group		
< 60	26,469	30.7
60-74	38,042	44.0
≥ 75	21,875	25.3

2.2 Statistical methods

Mortality related to IHD was estimated by computing the relative survival rate using the Hakulinen method, as the ratio of the observed to the expected rate. The observed survival rate for all causes of death was calculated by the Kaplan-Meier method based on the data of the NRKI. The expected survival rate was calculated from life tables which are freely available on the website of the Czech Statistical Office (CZSO, 2012). The log-rank test was used to assess differences between survival curves. Traditional age standardization was performed with weightings derived from the initial age structure of the study group.

3. Results

3.1 Observed and relative survival

Overall, 3,124 patients died within the first 30 days following their PCI (3.6 %); 10,388 other patients deceased during following 5 years. Figure 1 shows the observed survival (OS) and relative survival (RS) in men and women alive on day 31. OS was higher in men (82.7 % vs. 78.3 %), as well as RS (97.7 % vs. 94.2 %). In average, women were much older than men; thus age standardization would be appropriate. Age-standardized OS was higher in women (81.9 % vs. 80.4 %), age-standardized RS was still higher in men (98.5 % vs. 94.2 %) (Table 2).

This study showed significantly reduced relative long-term survival in women compared to men of all age groups. Patients older than 74 years were surviving better than younger ones (Figure 2). In men older than 75 years, the relative survival rate increased even to 103.5 % (Table 2).

Table 2. Crude and age-adjusted 5-year OS and RS rates of patient with survival > 30 days

	Observed survival (%)			Relative survival (%)		
Age group	Men	Women	All	Men	Women	All
< 60	92.5	93.4	92.6	97,2	95.6	96.9
60-74	82.3	83.2	82.6	96.5	91.2	94.7
≥ 75	62.5	65.6	64.0	103.5	97,9	100,6
Crude	82.7	78.3	81.3	97.7	94,2	96.6
Age-adjusted	80.4	81.9	81.0	98.5	94.2	96.9
Difference	(-2.3)	(+3.6)	(-0.3)	(+0.8)	(+0.0)	(+0.3)

3.2 Limitations of the study

In calculations of the expected survival in the study group it has to be assumed that survival in a general population is unaffected by deaths related to the disease of interest. If the prevalence of that condition in the general population is low enough, then this will have little impact (Nelson et al., 2008). Unfortunately, this is not the case because IHD is definitely not the rare disease, especially in the advanced age. Furthermore, patients indicated for PCI were selected with respect to their overall medical condition. Both reasons could have affected the calculated relative survival rates to the better results.

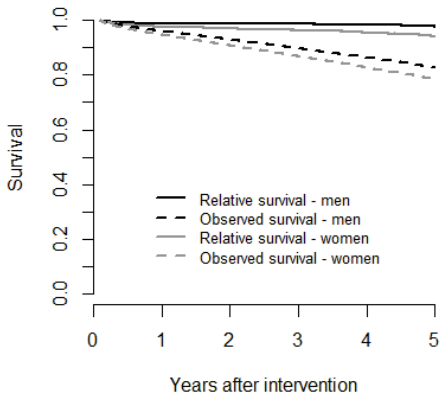


Figure 1. OS and RS curves following a first PCI by sex in patients with survival > 30 days

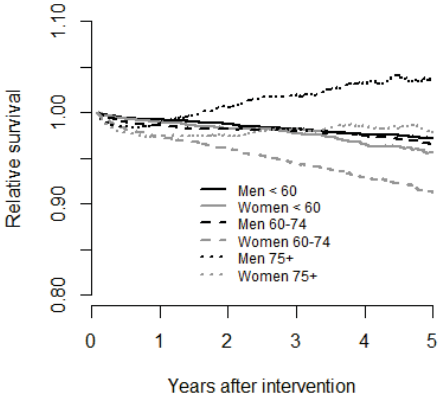


Figure 2. RS curves following a first PCI by age group and sex in patients with survival > 30 days

4. Conclusion

Age-standardized OS was higher in women while age-standardized RS was higher in men. It means that women after PCI deceased less than men but often in a consequence of IHD. Relative survival was lower in women and in patients below 75 years of age. This was most likely due to acceptance of patients with more comorbidity among the younger patients and/or the high prevalence of IHD in the general population.

5. References

Czech Statistical Office. 2012. Úmrtnostní tabulky za ČR od roku 1920 [online]. Praha: Czech Statistical Office, updated 2012-09-16 [cit. 2012-10-30]. URL: http://www.czso.cz/csu/redakce.nsf/i/umrtnostni_tabulky.

Nelson CP, Lambert PC, Squire IB, Jones DR. 2008. Relative survival: what can cardiovascular disease learn from cancer? European Heart Journal 29: 941-947.

The Statistical Office of the European Union. 2013. Causes of death - Absolute number (Annual data) [online]. Luxembourg: Unit F5: Health and food safety; Crime [cit. 2013-03-06]. URL: http://epp.eurostat.ec.europa.eu/portal/page/portal/health/public_health/data_public_health/database.

System of equine fitness evaluation based on time-frequency analysis

Igor Feigler[1], Michalis Zervakis[2], Jiří Holčík[1,3]

[1] *Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic; e-mail: 372231@mail.muni.cz*
[2] *Technical University of Crete, Greece; e-mail: michalis@display.tuc.gr*
[3] *Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia; e-mail: holcik@iba.muni.cz*

Abstract

In this paper we study the equine stress test ECG. We are using the time-frequency analysis in order to examine the changes in frequencies throughout the time interval of the experiment. The goal of this study is to describe the fitness state of the subjects according to information obtainable from ECG signals. For this purpose we have extracted certain features from the signal, namely power and Poincare plot descriptors. Cluster analysis is then used to create groups of features' values in order to interpret the heart's reaction to the stressing. We found, that it is possible to associate these groups with the testing protocol and thereby describe the state of stress.

Key words

ECG, health and fitness evaluation, equine stress test, time-frequency analysis

1. Introduction

The aim of this study is to use the equine stress test electrocardiogram (ECG) to create a system of fitness evaluation. The system is based on the time frequency analysis of signals and comparison among subjects. We extracted features from the time-frequency transform of the signal and these we analysed. The idea of the system was to associate the vectors of descriptors' values at certain frequency bands and time windows with the stress test protocol. These vectors were therefore classified by means of clustering.

Through this system we expect to have easier capturing of abnormalities either in health or in fitness of the subject than the plain time-frequency transformation result is able to provide. This might be very useful for quick and easy identification of problematic horses, or for example potential diseases of heart or any other that has effect on performance. Another use might also be the easy comparison of subjects leading to quick detection of heart adaptability, which is a key factor of performance. Therefore the possibility of using this system as a measure of racing ability might exist.

2. Dataset

The dataset includes complete equine ECG signals from fourteen subjects that were screened during a stress test. All these tests were done using a treadmill at a veterinary clinic and were supervised by veterinary surgeons.

3. Frequency domain analysis

As our system is based on time-frequency analysis of the signal, we needed to determine the length of the time window we would use. The frequency analysis helped us to decide this and also discovered that the highest frequency involved in the signal was much lower than original sampling rate. Therefore, we decimated the sampling frequency.

4. Time-frequency analysis

From all the time-frequency transform methods we chose to use the Fourier transform based short time Fourier transform (STFT). We chose to work with the hamming window and to overlap each successive window.

5. Feature extraction

The result of the STFT was only a step for us to be able to extract some features that would describe the signal. For that purpose we decided to deal with a simpler image of the power averages as it is much more suitable for classification and clustering. Therefore we divided the time axis into intervals of the 60 seconds length and the frequency axis into bands. It is also fact, that we only use frequency up to the value of 26.25Hz in here, since higher frequencies showed no visible amplitudes and were therefore considered as unimportant.

5.1. Power features

For each sub-matrix created by the bands we calculated the power. Power of a signal is defined as sum of the second powers of al+l the values divided by the number of these values. In order to get the same scale among all subjects we normalized the power values.

The next step was cluster analysis. We divided all the vectors of normalized power values for the specific frequency bands among subjects into clusters. We were trying to depict the real amount of stress that the subject felt during a particular minute. For this analysis we used the k-means clustering with *k* equal to three as this number turned out to be the best after few trials.

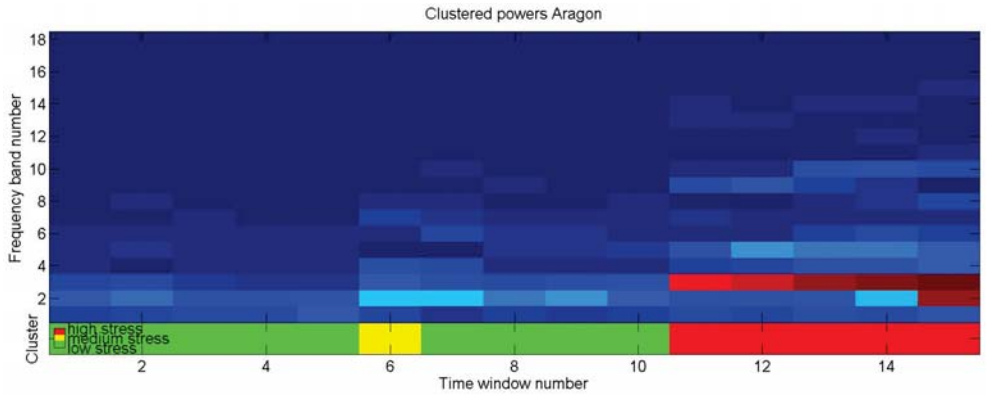


Figure 1. Visualisation of the power values extracted in specific time and frequency bands and its clustering for Aragorn

Figure 1 shows a sample visualisation of the power analysis result along with the clustering result. The upper part of Figure 1 is a result of power analysis. On the x axis there are time bands; y axis represents the frequency bands. The colour stands for the normalized power value. The bottom part of this image represents the cluster in which each vector of powers belongs.

Our image as we can see was not perfect. The stage power changes are very sudden causing wrong clustering in some cases. Therefore we decided to overlap our time windows.

The first visualisation of all the subjects' results is presented in Figure 2. The picture is constructed with an upper part, which represents the power of the signal for time and frequency windows. The lower part then represents the cluster, into which each vector of powers in specific frequency bands for a time window belongs. White lines across the whole picture represent the border between two horses. The *x* axis represents the time window number, which cumulates with the subjects.

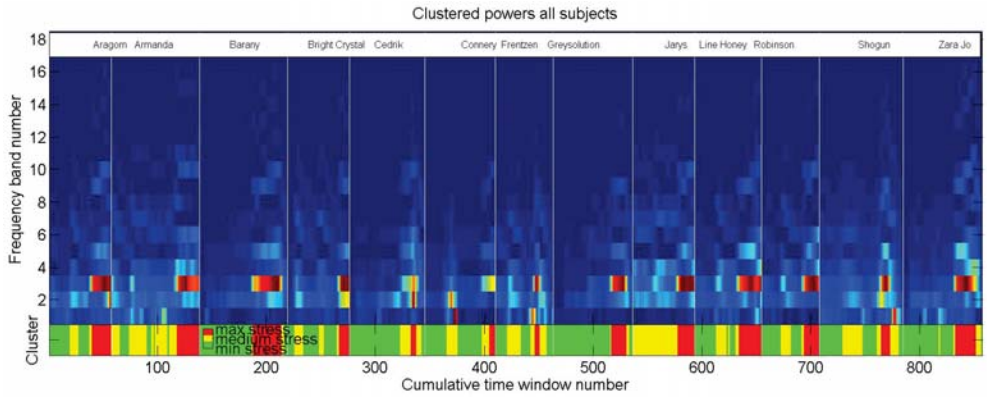


Figure 2. Visualisation of the power values extracted in specific frequency bands and overlapping time intervals and its clustering for all subjects

5.2. Poincare Features

Next feature we extracted from our signal represents the geometric domain. We used these descriptors of the Poincare plot: the standard deviation in the direction of the identity line (called SD2) and the standard deviation in the direction orthogonal to the identity line (called SD1) (Bravi et al., 2011). Poincare plot represents the display of a generic sample *n* of the time series and as a function of the sample *n-1* (Linn et al., 2010). As stated, one creates a Poincare plot from a signal in time domain. As we did a great analysis in the frequency plane, we can easily filter the signal, so that we would use only the frequencies considered to carry the useful information. We decided these frequencies to originate in the first five frequency bands. That means in this place we use only frequencies up to 6.75 Hz. As we had a matrix of these measures for all horses, we again needed to apply normalization in order to have the same scale.

Figure 3 shows the resulting clustered normalized Poincare descriptors for all horses. The upper part shows the dynamics of the descriptor values in time and the lower part shows the cluster into which each time window belongs. Subjects are separated from each other by thin red lines. The *x* axis represents the time window number, which cumulates with the subjects.

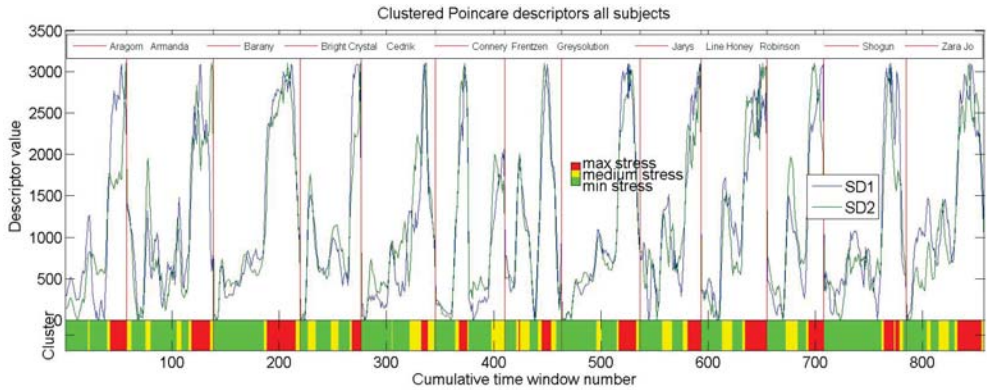


Figure 3. Visualisation of the Poincare descriptors' values extracted in specific time and frequency bands and its one-stage three-means clustering for all subjects

5.3. Combined Feature visualisation

Now, as we have these three clusters defined by the two Poincare descriptors, we can add them to the clustered power image to capture similarities.

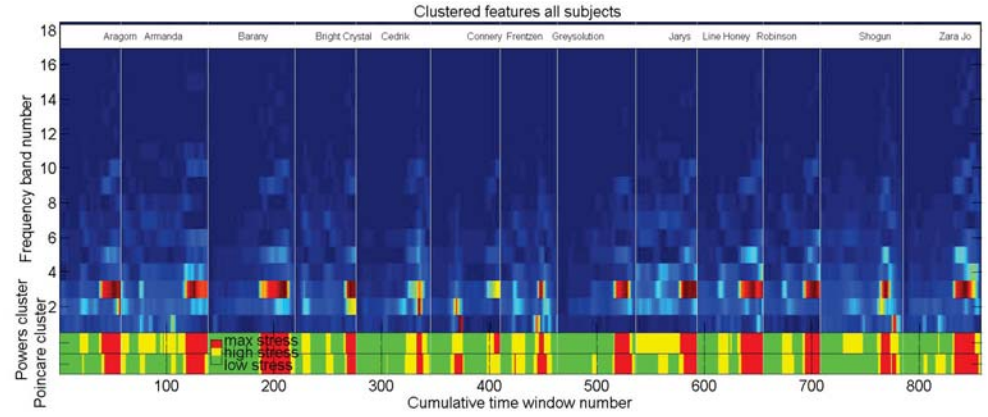


Figure 4. Visualisation of the power values extracted in specific frequency bands and overlapping time intervals and its one-stage three-means clustering with the result of the clustering of the Poincare descriptors values for all subjects

Figure 4 shows the complete results we accomplished for overlapping time windows. The upper part represents the normalized power of the signal; the lower part consists of two cluster analyses. The upper cluster analysis represents the powers cluster; the lower cluster analysis represents the Poincare descriptors cluster. Subjects are always separated with a thin white line. The *x* axis represents the time window number, which cumulates with the subjects.

6. Discussion

This study has several limitations. Firstly, the dataset is not large enough to provide statistics of differences among subjects and by that to prove our system provides contributive results. Secondly, only few features were extracted to describe the ECG dynamics. A possible expansion of this work is the use of other features, possibly describing different domains, such as the invariant, or the statistic. Also, a combination of used, or other features might describe the ECG variability better than each separately. Additionally, one could compare to our assumptions about the fitness state of horses based on the ECG features to some actual physiological measures, such as lactate rates during the experiment. Other comparison might be done with the handicap information of the subjects.

7. Conclusion

In this study we analysed the equine ECG signals recorded during a stress test. We created a system helping us to find similarities and differences in responses to the different phases of stressing.

8. References

Bravi A., Longtin A., Seely A.J.E., 2011. Review and classification of variability analysis techniques with clinical applications. BioMedical Engineering OnLine 2011 10:90.

Lin C.W., Wang J.S., Chung P.C., 2010. Mining Physiological Conditions from Heart Rate Variability Analysis. IEEE Computational Intelligence Magazine 5:50-58.

Statistical evaluation of recurrent events in chronic myeloid leukaemia

Petra Kovalčíková[1], Tomáš Pavlík[2], Eva Janoušová[2]

[1] *Faculty of Science, Masaryk University, Brno;*
e-mail: kovalcikova@mail.muni.cz
[2] *Institute of Biostatistics and Analyses, Masaryk University, Brno*

Abstract

This work aims to introduce and apply the methodical background for statistical evaluation of recurrent events in chronic myeloid leukaemia. A non-parametric approach is adopted that is based on standard methods of survival analysis and analysis of competing risks. Moreover, an approach based on the so-called multi-state models can be used. In the application on real data sets, the results of analyses of multiple remission periods in patients receiving imatinib for chronic myeloid leukaemia are presented.

Key words

Stochastic modelling, competing risks, survival analysis, recurrent events

1. Introduction

In many clinical studies in which death is not the event of interest, subjects may experience the so-called recurrent event, i.e., a pre-defined event that may occur repeatedly several times during the follow-up period. As an example of recurrent event data, multiple periods of relapses or remissions in chronic myeloid leukaemia (CML), rheumatoid arthritis, breast cancer, or re-hospitalizations of patients after recurrent heart attacks or vascular brain strokes can be mentioned. The aim of this work is to present and apply estimates of the so-called current survival measures in CML patients.

2. Methods

Non-parametric statistical methods were used for estimating two principal characteristics of the current CML treatment: the probability of being alive and leukaemia-free in time after CML therapy initiation, denoted as the current cumulative incidence of leukaemia-free patients (*CCI*); and the probability that a patient is alive and in any leukaemia-free period in time after achieving the first leukaemia-free period on the CML treatment, denoted as the current leukaemia-free survival (*CLFS*). Being leukaemia-free was defined as being in the complete cytogenetic remission (CCgR), which is defined as the complete eradication of karyotypically apparent Philadelphia chromosome positive metaphases.

The *CCI* can be written using the common cumulative incidence functions corresponding to the achievements of CCgR, $I_i(t)$, losses of CCgR, $I_i^*(t)$, or death after ith achievement of CCgR, $I_i^{**}(t)$, which can be estimated using the standard Aalen-Johansen estimator (Marubini and Valsecchi, 2004):

$$CCI(t) = \sum_{i=1}^{r} I_i(t) - \sum_{i=1}^{r} I_i^*(t) - \sum_{i=1}^{r} I_i^{**}(t) = \sum_{i=1}^{r} [I_i(t) - I_i^*(t) - I_i^{**}(t)]. \quad (1)$$

Moreover, the *CLFS* can be written using the survival functions, where the event of interest is death in the *i*th CCgR or *i*th loss of CCgR, $S_i^*(t)$, or *i*th achievement of CCgR or death prior it, $S_i(t)$. These survival functions can be easily derived with the standard Kaplan-Meier estimator (Kaplan and Meier, 1958):

$$CLFS(t) = S_1^*(t) + \sum_{i=2}^{r} [S_i^*(t) - S_i(t)]. \quad (2)$$

The methods for estimating *CCI* and *CLFS* curves were described in more details elsewhere (Pavlík et al., 2011). Moreover, the current survival measures were compared with the common ways of patient outcome assessment, common leukaemia-free survival (*LFS*) and cumulative incidence (*CI*), which are, however, not well suited for quantification of CML treatment outcomes, because these measures cannot account for multiple disease remissions that can be achieved using sequential therapy in CML. R software package currentSurvival was used for the estimation.

3. Results

In total, 723 Czech CML patients in chronic phase, who received first-line imatinib between July 2003 and December 2011 and who were registered in the Czech databases CAMELIA (http://www.camelia.registry) and INFINITY (http://www.leukemia-cell.org/en/databaze/), were used for the analysis.

In Figure 1, the left side graph represents the resulting estimates of the common *CI* and the *CCI* curve as well as the 95% point-wise bootstrap confidence intervals. Furthermore, the right side graph shows the resulting estimates of the *CLFS* and the common *LFS* curves. These estimates are also accompanied with the 95% point-wise bootstrap confidence intervals.

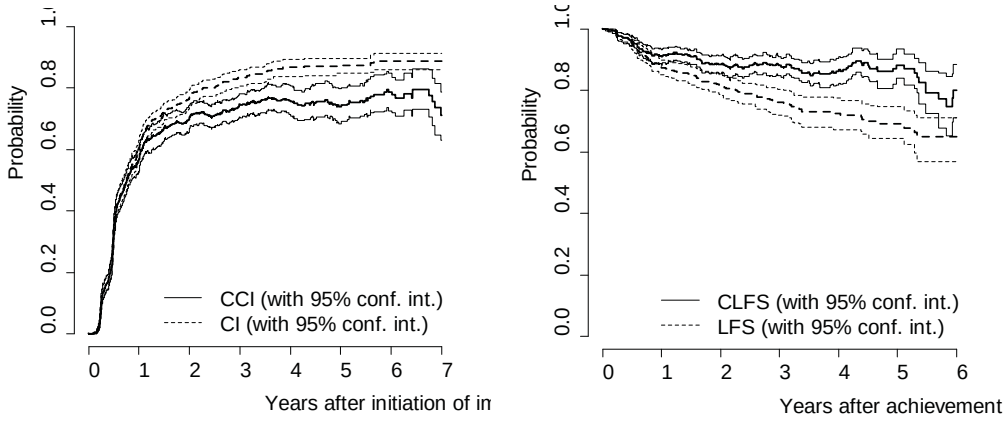


Figure 1. *CCI*, *CI* (left) *CLFS*, and *LFS* (right) estimates for 723 Czech CML patients in chronic phase, who received the first-line imatinib between July 2003 and December 2011

Regarding all 723 patients, the estimated *CCI* at 3 and 5 years after starting imatinib therapy was 74.4% (95% CI: 70.3%–78.0%) and 73.8% (95% CI: 68.4%-79.7%), respectively. On the other hand, the common *CI* at 3 and 5 years after starting imatinib was estimated as 83.5% (95% CI: 80.7%–85.9%) and 87.4% (95% CI: 84.8%-89.6%), respectively. Thus, the estimated difference between the *CCI* and *CI* curves reached 9.1% and 13.6% at 3 and 5 years after starting imatinib, respectively.

Only 553 patients (76.5%) who achieved at least one CCgR were available for the *CLFS* calculation. The estimated *CLFS* at 3 and 5 years after achieving the first CCgR was 88.0% (95% CI: 84.8%-90.9%) and 88.2% (95% CI: 83.8%-93.3%), respectively. The *LFS* was estimated as 76.2% (95% CI: 72.0%-80.3%) and 69.1% (95% CI: 64.4%-74.8%) at 3 and 5 years after achieving the first CCgR, respectively. Therefore, at 3 and 5 years after the achievement of the first CCgR, the difference between the *CLFS* and *LFS* estimates reached 11.8% and 19.1%, respectively.

4. Conclusions

The common *CI* overestimates the probability of being alive and in CCgR after starting first-line imatinib therapy, whereas the common *LFS* underestimates the probability of being alive and in CCgR after the achievement of first CCgR on imatinib. Thus, both current survival measures, the *CCI* and *CLFS*, more reliably illustrate a CML patient's disease status in time because they account for multiple leukaemia-free periods during the treatment course.

5. References

Kaplan EL, Meier P. 1958. Nonparametric estimation from incomplete observations. Journal of American Statistical Association, 58, 457-481.

Marubini E, Valsecchi MG. 2004. Analysing Survival Data from Clinical Trials and Observational Studies. John Wiley & Sons.

Pavlík T, Janoušová E, Pospíšil Z, Mužík J, Žáčková D, Ráčil Z, Klamová H, Cetkovský P, Trněný M, Mayer J, Dušek L. 2011. Estimation of current cumulative incidence of leukaemia-free patients and current leukaemia-free survival in chronic myeloid leukaemia in the era of modern pharmacotherapy. BMC Medical Research Methodology, 11:140.

Risk factors for rehospitalization and mortality for cardiovascular event in a consecutive group of patients after first hospitalization for acute heart failure

Michal Svoboda

IBA & RECETOX, Faculty of Science, Masaryk University, Brno;
e-mail: 379969@mail.muni.cz

Abstract

Acute heart failure is one of the most common causes of death in the developed countries. It is a condition with high risk of hospitalization mortality and mortality in medium time. There are many studies that analyzed mortality risk factors, but only one study, the COACH study, analyzed the risk factors for rehospitalization. Individual rehospitalizations are expensive and reduce the quality of life. AHEAD database was used as a data source, namely a consecutive subset of 608 patients. For analysis of risk factors it is necessary to use multistate survival models, which are models with a final number of states, in which the patients can enter during the follow-up period. Using multistate survival models we are able to determine which factors affect individual transitions. Peripheral vascular disease (PVD) was the most important risk factor in patients, who were not rehospitalized. Patients with PVD had a 3.7 times higher risk of death than healthy patients. In patients, who were rehospitalized, aortic stenosis was analyzed as the most important risk factor of death. Individuals with aortic stenosis had a 3.1 times higher risk of death than others. Aortic stenosis was analyzed as the most important risk factor for rehospitalization, too. Patients with aortic stenosis had a 1.9 times higher risk of being hospitalized.

Key words

acute heart failure, risk factor, rehospitalisation, mortality, multistate survival model, hazard ratio

1. Introduction – heart failure

Heart failure is a state in which the heart is unable to fulfil its function of pump. This means that does not go enough blood into the circulatory system. We can distinguish between left and right heart failure. If left part of the heart is damaged, then it does not draw enough blood to organs. With right heart failure there is not enough blood in the lungs. Further we can divide heart failure on acute and chronic. Acute heart failure (AHF) occurs suddenly in a relatively healthy heart, while we can talk about chronic heart failure, when the failure recurs.

1.1. Causes of AHF

The most common cause of AHF is cardiomyopathy, which is a summary term for all damage to the myocardium. The other reasons of heart failure are heart attack, heart arrhythmias, hypertensive crisis or cardial tamponade.

1.2. Frequency of AHF

AHF occurs in 0.4–2% of the total population and in the central Europe up to 1.3% of the local population suffer from this problem (Špinar a Vítovec, 2007). Its frequency increases with age. Every eleventh individual at age 80–90 years suffers from AHF.

1.3. Prognosis after AHF

Despite progress in treatment, prognosis for patients AHF is poor. This problem is caused by the fact that diagnosis and treatment of acute heart failure are very medically and economically demanding. About 70% of patients die within 5 years from heart failure (Postmus et al., 2011), of which 25% of individuals die within the first year. AHF influences other rehospitalization from cardiovascular causes. Around 45% of patients hospitalized with heart failure are rehospitalized during the 12 months and the risk of death then increases up to 60%.

2. Methods

2.1. Statistical methods

Survival analysis is a set of the statistical methods, by which we can analyze the time to occurrence of observed events. The analysis is characterized by two functions – survival function and risk function.

The Cox regression model is a statistical method, by which we can determine the relationship between patient's survival and the explanatory variables.

The hazard ratio (HR) for the two groups (e.g., diabetic and non-diabetic) provides information on how many times one group has a higher risk of occurrence of the event than the other one.

2.2. Patients

Data source is the AHEAD database, which was established in 2006 and was terminated in 2012 with more than 8,600 patient records after AHF. Analysis of risk factors was performed only on a subset of 608 consecutive patients from the University Hospital Brno. Men slightly prevailed in this cohort (53%). The average age of the dataset was 72.1 years, with 61.2% of subjects older than 70 years. Representation of other factors, which were analyzed as risk factors in individual studies, is shown in Table 1.

3. Results

2.1. Risk factors of mortality

Of all 608 patients, 487 were not rehospitalized. Of these patients, 219 eventually died. The most important risk factor for mortality in patients without rehospitalization was peripheral vascular disease (PVD). Patients with PVD have a 3.7 times higher risk of death than healthy individuals (Figure 1). 121 patients were rehospitalized. Of these patients, 104 returned home and eventually 40 died. The most important risk factor of death in these patients was aortic stenosis. Individuals with aortic stenosis have a 3.1 times higher risk of death than others (Figure 2).

Table 1. Representation of risk factors from studies (n=608)

	Number of patients	%
Sex		
Male	322	53.0
Female	286	47.0
Age > 70 years	372	61.2
Ejection fraction (EF) < 40	281	46.2
Uric Acid (> 420 umol/l for M, 120 g/l for F)	259	42.9
Diabetes mellitus	249	41.0
Anaemia (Hb < 130 g/l for M, 120 g/l for F)	202	33.8
Heart attack	195	32.1
Pulmonary edema	118	19.4
Hypertensive crisis	55	9.0
Hyperkalemia (K > 5.5 mmol/l)	27	4.4
Renal failure	14	2.3
Hyponatraemia (Na < 130 mmol/l)	26	4.3

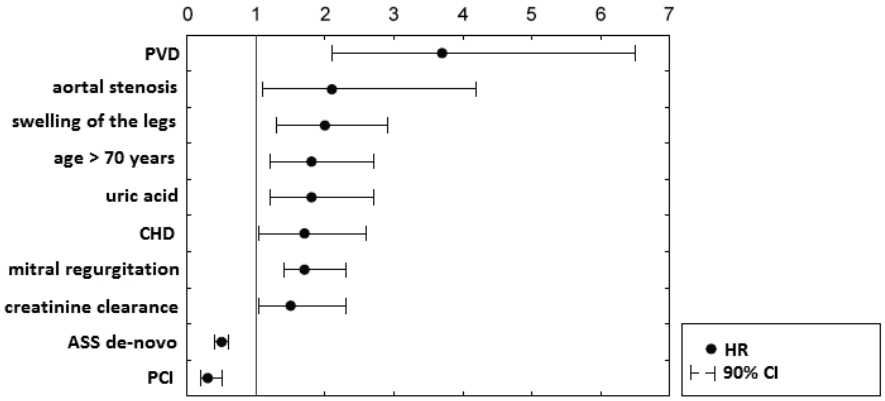


Figure 1. Mortality risk factors in patients without rehospitalisation (n=487)

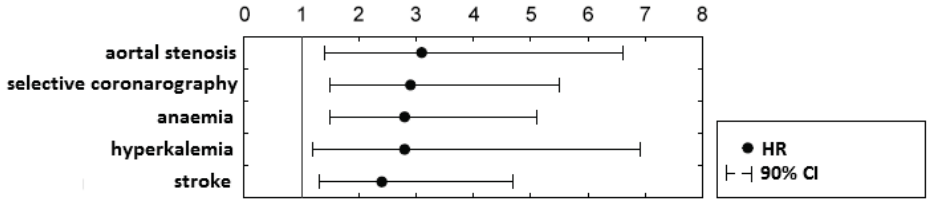


Figure 2. Mortality risk factors in patients after 1st rehospitalisation (n=104)

2.2. Risk factors of 1st rehospitalization

Of all 608 patients, 121 were rehospitalized. Aortic stenosis was the most important risk factor for rehospitalization as well as for death in patients after rehospitalization. Patients with narrowed aortic valve have a 1.9 times higher risk of being rehospitalized (Figure 3).

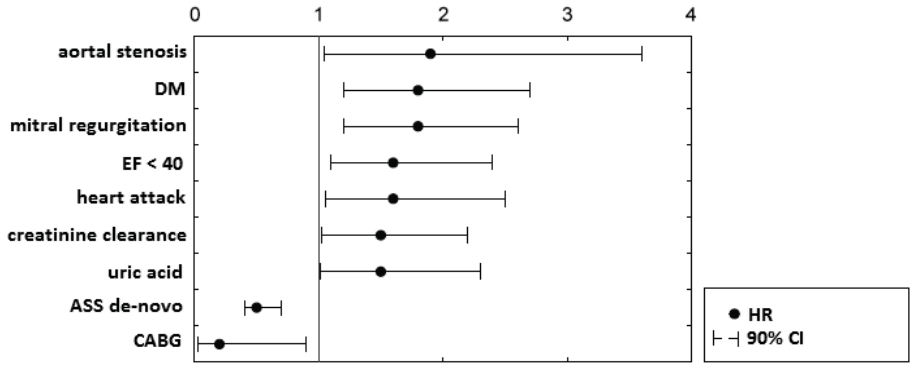


Figure 2. Risk factors for 1st rehospitalization (n=608)

3. Conclusion

It was found that aortic stenosis became a factor increasing the risk of death and rehospitalization. Patients with mitral regurgitation have an increased risk of death and rehospitalization, too. Regarding the biochemical and chemical parameters, low value of creatinine clearance, high uric acid, anemia and hyperkalemia were analyzed as risk factors for death. Creatinine clearance and uric acid were also demonstrated as the factors increasing the risk of rehospitalization. Conversely PCI and bypass, surgery solving heart attack, reduce the risk of death and rehospitalisation.

4. References

Postmus D, van Veldhuisen DJ, Jaarsma T, Luttik ML, Lassus J, Mebazaa A, Nieminen MS, Harjola VP, Lewsey J, Buskens E, Hillege HL. The COACH risk engine: a multistate model for predicting survival and hospitalization in patients with heart failure. 2012. European Journal of Heart Failure 14: 168-75.

Špinar J, Vítovec J. Jak dobře žít s nemocným srdcem. Praha: Grada Publishing, 2007. 255 p. ISBN 978-802-4718-224.

Proceedings of the 9th Summer School on Computational Biology
Stochastic Modelling in Epidemiology

Editors: Tomáš Pavlík, Ondřej Májek
Cover: Radim Šustr

Published by Masaryk University
www.muni.cz

Printed by Tiskárna KNOPP, s.r.o., Nádražní 219, 549 01 Nové Město nad Metují
1st edition, 2013
100 copies

ISBN 978-80-210-6305-1