MASARYK UNIVERSITY
Faculty of Science
Department of Mathematics and Statistics

DISSERTATION

Tomáš Pavlík

Brno 2010

MASARYK UNIVERSITY
Faculty of Science
Department of Mathematics and Statistics

Tomáš Pavlík

MATHEMATICAL MODELS IN ONCOLOGY

Dissertation

Supervisor: prof. RNDr. Ivana Horová, CSc. Brno 2010

Bibliographic entry

Author	Tomáš Pavlík
Title of dissertation	Mathematical models in oncology
Název dizertační práce	Matematické modely v onkologii
Study programme	Mathematics
Study field	Probability, statistics and mathematical modelling
Supervisor	prof. RNDr. Ivana Horová, CSc.
Year	2010
Keywords	cancer; cancer survival benchmarks; relative survival; Cox model; frailty; regression diagnostics
Klíčová slova	maligní onemocnění; referenční hodnoty přežití; Coxův model; frailty; regresní diagnostika

Acknowledgements

I would like to thank to my supervisor prof. Ivana Horová for all her help she has given me during my study. I am also grateful for cooperation with doc. Ladislav Dušek for his assistance and mentorship in the field of data analysis. I would also like to thank the Institute of Health Information and Statistics of the Czech Republic as well as the Czech and Slovak hemato-oncology centres involved in the Camelia project for providing me with data for my work. My special thanks go to my colleagues and friends, Ondřej Pokora and Ondřej Májek, for their helpful advices and reading through my draft copies. Ondřej Májek has also significantly contributed to Chapter 3 as the author of model for cancer incidence. Additionally, I also highly appreciate the friendship with other members of our research group, especially Jan Mužík and Eva Janoušová, with whom I worked closely and puzzled over many of the same problems. The last but surely not least, I am also grateful for all the support I have received from my family whilst researching and writing up this dissertation.

Abstract

The emphasis of this dissertation lies on mathematical methods for modelling of cancer patient data. First chapter gives an introduction to cancer epidemiology and definition of basic terms needed in following chapters. Moreover, some key issues for the population-based cancer data assessment are discussed. Chapter 2 is a methodical chapter on mathematical methods used for cancer survival assessment. Some basic survival analysis methods are recalled and the concept of relative survival is introduced. Then, Cox proportional hazards model is defined as the most popular model for cancer survival modelling, and the estimation principles of regression coefficients using the partial log likelihood together with the methods for regression diagnostics are given. Furthermore, frailty models which can be considered as an extension of the Cox model are introduced as well as mixture cure fraction models for the modelling of relative survival. In Chapter 3, the emphasis lies on the estimation of prevalence of patients requiring active anti-tumour therapy that is accessible from population-based cancer registry data. The model is further applied on colorectal cancer data from the Czech National Cancer Registry to model the number of potentially treated patients with colorectal carcinoma in the Czech Republic in 2011. Chapter 4 presents the survival benchmarks for Czech cancer patients, and provides an overview of the survival rates achieved since 1990. Moreover, a comparison is provided between the Czech and global data concerning the five-year relative survival rates in selected cancer diagnoses. Finally, Chapter 5 presents a Cox regression model for the achievement of the complete cytogenetic or major molecular response to a modern targeted therapy in Czech and Slovak patients in chronic phase of chronic myeloid leukemia.

Abstrakt

Tato práce je věnována problematice modelování dat pacientů s maligním onemocněním pomocí matematických metod. V první části práce jsou definovány hlavní epidemiologické charakteristiky používané v dalších kapitolách. Dále jsou v úvodní kapitole uvedeny některé klíčové prvky hodnocení populačních onkologických dat. Kapitola 2 je věnována matematickým metodám pro hodnocení přežití onkologických pacientů. Vedle základních metod analýzy přežití je zaveden pojem relativní přežití a jsou definovány metody jeho odhadu. Je definován tak Coxův model proporcionálních rizik spolu s metodou odhadu regresních koeficientů a metodami regresní diagnostiky. Dále je v kapitole 2 uvažováno rozšíření Coxova modelu ve formě tzv. *frailty* modelů, a tzv. *mixture cure fraction* model pro modelování relativního přežití. Kapitola 3 představuje model pro odhad počtu onkologických pacientů potenciálně léčených protinádorovou léčbou, který je aplikován na data českého Národního onkologického registru s cílem odhadnout počet protinádorově léčených pacientů s kolorektálním karcinomem v ČR v roce 2011. Kapitola 4 prezentuje referenční hodnoty přežití pro hodnocení výsledků léčebné péče o onkologické pacienty v ČR a zpřehledňuje vývoj dosahovaných hodnot přežití od roku 1990. U vybraných onkologických diagnóz kapitola nabízí srovnání přežití dosahovaného v ČR s mezinárodními daty. Nakonec, kapitola 5 je věnována modelování doby do dosažení kompletní cytogenetické nebo významné molekulární odpovědi na léčbu moderní farmakoterapií u pacientů v chronické fázi chronické myeloidní leukémie pomocí Coxova modelu proporcionálních rizik.

Table of contents

Introduction to cancer epidemiology

1

Cancer is without any question one of the major health issues worldwide even though the clinicians and scientists has been fighting against it for decades. Moreover, besides that fact that cancer represents a significant burden of disease, it also represents an important socioeconomic factor affecting the society. The number of newly diagnosed patients is increasing in both the developed and the developing world due to increased life expectancy, urbanization, environmental pollution and adoption of unhealthy lifestyle. Although progress has been achieved in western countries with prevention programmes and improvements in cancer detection and medical care, the burden of cancer is supposed to grow also in the future, mainly due to population ageing.

The Czech Republic is no exception in this respect as it ranks among countries with the highest cancer load worldwide with tens of thousands of new cancer patients being newly diagnosed every year [87]. Moreover, there are hundreds of thousands of cancer patients who were diagnosed and treated in previous years, constituting a indispensable burden of the Czech health care system. From the financial perspective, cancer is associated with the economic cost represented with expenditures on cancer prevention programmes, screening and treatment, the economic cost represented with time and effort spent by patients and their relatives and the economic cost represented with lost productivity due to cancer-related disability and premature death.

Cancer epidemiology can be defined as the study of the distribution and determinants of cancer in specified populations, and the application of this study to control of health problems [26]. From the biostatistician's perspective, the principal aim of cancer epidemiology should be to analyse population-based data to enable for drawing conclusions about the time trends and the risk levels associated with different groups of individuals. The key idea that we have to be able to distinguish between statistically significant trends and random fluctuations is especially true in analysis of epidemiological trends. As already mentioned, cancer epidemiology is focused on populations rather than on separate individuals; thus the typical questions which the epidemiologists can ask considering cancer can be as follows:

- How does the number of newly diagnosed patients change over time?

- How much does the number of new patients vary from place to place?
- How are the patients likely to survive 5 years being diagnosed with cancer?
- How many patients ever diagnosed with cancer will be alive next year?

Population-based assessment of cancer epidemiology is vital for functional monitoring of health care system in the widest sense, i.e. for the evaluation of all of its components. However, the main goals in the fight against cancer to which the analyses of epidemiological characteristics should contribute can be stated as follows:

- Lowering number of new cancer patients and increasing their chance not to die of cancer.
- Improving quality of life of cancer patients.
- Making the best use of available resources for cancer diagnosis and treatment.

Fortunately, monitoring of health care quality has become an integral part of Czech cancer care where data are collected and analysed using the population-based registries or clinical studies. Decisions in health care management, however, must be based on carefully selected parameters and analyses wherein the value of information and the precision of interpretation are incontestable. This requires objective and well defined measures of risk in order to make relevant comparisons between two conditions of interest. These measures might take the form of the probability of being diagnosed with cancer or of dying from it, or they might consist of the survival rate or the probability of cancer recurrence. In all cases, these measures are most often calculated as a ratio between the number of observed events and the number of individuals at risk within a given period of time.

1.1 Measures of cancer occurrence

There are two main measures of cancer frequency, called incidence and prevalence, commonly introduced in epidemiology literature. First of them, cancer incidence, refers to new cancer cases occurring among whole population of individuals that are at risk. It can be expressed as the overall number of newly diagnosed patients, however, usually it is expressed as the number of cancers per 100,000 people at risk [14]. Main reason for expressing the cancer incidence

per 100,000 people is the comparability over different populations. Thus we can formulate the cancer incidence as follows

$$\text{Incidence rate} = \frac{\text{New cancers}}{\text{Population}} \times 100,000. \tag{1.1}$$

The number of newly diagnosed cancer patients, represented by the numerator, may include multiple primary cancers occurring in one patient because, obviously, each patient can suffer from more than one type of cancer. On the other hand, the incidence rate would not in general include cancer recurrences, i.e. return of the cancer in the same place where the cancer first originated. Most frequently, the annual incidence rates are calculated, referring to the number of cancer cases newly diagnosed during the particular calendar year. Cancer incidence is the main characteristic of cancer dynamics in a given population as it can flexibly reflect improvements in cancer care and prevention, namely in prevention programmes, organized screening programmes, and improvements in diagnostic methods.

Cancer prevalence is defined as the proportion of patients with present or past diagnosis of cancer alive in a population at a specified time point. More specifically, this definition should be denoted as point prevalence for the numerator of the proportion comprises all patients who are alive and have the disease at that instant, irrespective of whether it was diagnosed recently or many years ago. Thus, cancer prevalence includes both newly diagnosed patients (cancer incidence) and patients diagnosed in the past, i.e. prevalence is a function of both past incidence and survival. In epidemiology, however, also the so-called period prevalence is defined, referring to the proportion of the population with cancer over a specific time period. For example, 2007 cancer prevalence can be calculated as the number of individual patients in a population that had a cancer in 2007. Like the incidence rate, also the prevalence can be alternatively expressed as the overall count or as the number of patients per 100,000 people alive in a population at a specified point in time. Cancer prevalence cannot be easily nor directly used for cancer dynamics assessment like the incidence rate, however, prevalence is fully representative of the overall cancer burden in a population of interest.

When talking about cancer prevalence, the epidemiologists mainly think of the so-called complete prevalence representing all patients ever diagnosed with cancer and alive at a given time point. However, complete prevalence is often not available from population-based data as the registration of cancer cases has not been working long enough yet. In this case, we are restricted to the so-called limited-duration prevalence that represents the number of patients alive at a specified time point diagnosed with cancer within the past x years. Limited-duration

prevalence is easily estimable from population-based data, e.g. by the counting method [14], and can be further adjusted for missing cancer cases using completeness indices [15].

Another epidemiology characteristic inherently associated with cancer is the mortality rate which represents the number of deaths, with cancer recorded as the cause of death, occurring in a population within a time period of interest, again usually expressed as the number of patients deceased due to cancer per 100,000 people at risk. We can then formulate the cancer mortality as follows

$$\text{Mortality rate} = \frac{\text{Death due to cancer}}{\text{Population}} \times 100,000. \tag{1.2}$$

Characteristic closely related to cancer mortality is cancer patient survival which is often used as a measure of cancer patient care. Patient survival rates represent a key parameter in oncology, and are routinely used both in the assessment of clinical experiments and in the analysis of cancer burden in the population [6, 18]. There are two main reasons for the estimation of cancer survival rates:

(i) We want to describe the outcome of patients diagnosed with cancer to assess the associated mortality. Such information can be further used for proposing and monitoring public health priorities. Moreover, the results can also be used for providing prognostic information for a newly diagnosed cancer patients.

(ii) We want to study mortality of different groups of cancer patients, i.e. we want to identify prognostic factors associated with varying survival of cancer patients.

However, the survival analysis on population level can lead from various reasons to spurious results. Therefore, the obtained results must be assessed very carefully, as survival rates are indicators of very complex population relations and trends, and improvements in patient survival do not necessarily result from a more effective treatment. On population level, higher survival rates may be associated with better diagnostic methods which make it possible to detect less advanced stages and to achieve better treatment results [98, 25].

1.2 Key issues in cancer epidemiology

There are several issues crucial for the population-based cancer data assessment, not solely associated with the statistical methods, that should be addressed prior to the definition of the methodological background.

1.2.1 Relative survival

Observed survival (overall survival in clinical studies) is the most basic measure of the survival experience of subjects under the study. However, its use in population-based studies can be misleading for the observed survival integrates all causes of death irrespective of the association with the disease of interest. Therefore, when the aim is to quantify the mortality due to the cancer under study, an approach with adjustment for the other causes need to be adopted. There are two standard methods applied for estimating the survival associated with one particular cause of failure in the population-based analyses, the method of cause-specific survival and the method of relative survival. The estimation of cause-specific survival proportions requires that reliably coded information on cause of death is available in data which is often problematic in population-based registries, because even if the record on cause of death is available, it can be difficult to determine whether or not cancer was the primary cause of death. The calculation of relative survival does not require precise information on the cause of death and can be easily obtained using population life-tables. It is for this reason that relative survival became a standard method for population-based survival analysis.

1.2.2 Calculation of specific rates

An important issue associated with the identification of the measures mentioned above is the calculation of the so-called specific rates. Obviously, as cancer rates vary widely with respect to many clinical and demographic factors, comparison of raw results could be misleading and standardisation for possible confounders is essential for making inference about any identified pattern. Above all, extent of disease, age, sex and time period of diagnosis are the most influential clinical and demographic variables that are needed to be accounted for in the population-based analyses. The extent of disease, expressed in oncology dominantly in form of the clinical stage, is mentioned first for the clinical stage is by means of patients life-expectation and anticipated financial budget impact of treatment the most influencing factor, even more influencing than age at diagnosis. An example of clinical stage influencing patient survival can be seen on Figure 1.1 which shows five-year observed and relative cancer survival rates in the Czech Republic in 2003–2005 period according to clinical stage in selected malignant neoplasms (MN). Age is another example of factor that is greatly influential with respect to the epidemiological characteristics mentioned above. In general, an age-adjusted rate is a weighted average of the age-specific rates, where the weights are the proportions of persons in the corresponding age groups of a standard population. The potential confounding effect of age is thus reduced when comparing

age-adjusted rates computed using the same standard population.

1.2.3 Data quality

Complete and relevant data are a keystone of every statistical model or analysis, but this is twice as true in epidemiology research. Obviously, the better the quality of data of a cancer or clinical registry, the better the possibilities for effective use of these data in planning and research. However, sometimes data completeness and relevancy are not enough for proper population-based analysis. In epidemiology studies in general, sufficiently long time series are required for the indication from epidemiological data, i.e. representative long-term profiles of newly diagnosed cancer cases, prevalence and mortality and a very good awareness of most important risk factors in population dynamics.

The Czech Republic disposes of a database of cancer patients which was established in 1976 and has been regularly updated and maintained since then (the Czech National Cancer Registry, CNCR). Data from the Reports on Malignant Neoplasms are assigned to unique birth certificate numbers of individual cancer patients, and are subsequently completed with records from the Follow-up Reports on Malignant Neoplasms. Keeping records on cancer cases within the CNCR has become an indispensable part of comprehensive cancer care; the CNCR has become a nationwide registry with 100% coverage of the Czech population, containing over 1.5 million records from the period 1976-2007. The CNCR contains patient data, data about the tumour and its clinical diagnostics, data on patient treatment and data about the patient after such treatment.

1.2.4 Different approaches to population-based survival rate estimation

Assessment of the survival rates is always performed retrospectively after a required follow-up period is reached; several different methods of time selection of patients can be found in the literature [9, 25, 80]. Logically, different methods lead to more or less different survival rates; therefore, the method of choice must be carefully considered. The so-called cohort analysis is considered to be the standard method, having been used in many comparative studies on cancer patients' survival in Europe, such as the EUROCARE [6]. The estimate of x-year survival rate in this method is based on the analysis of patients' records whose x-year follow-up after diagnosis ended, i.e. who had been diagnosed at least x years before the population-based database was closed. For example, in order to estimate the five-year survival for the period 2000–2005 by the cohort analysis, one would have to work with the cohort of patients diagnosed in 1995–2000

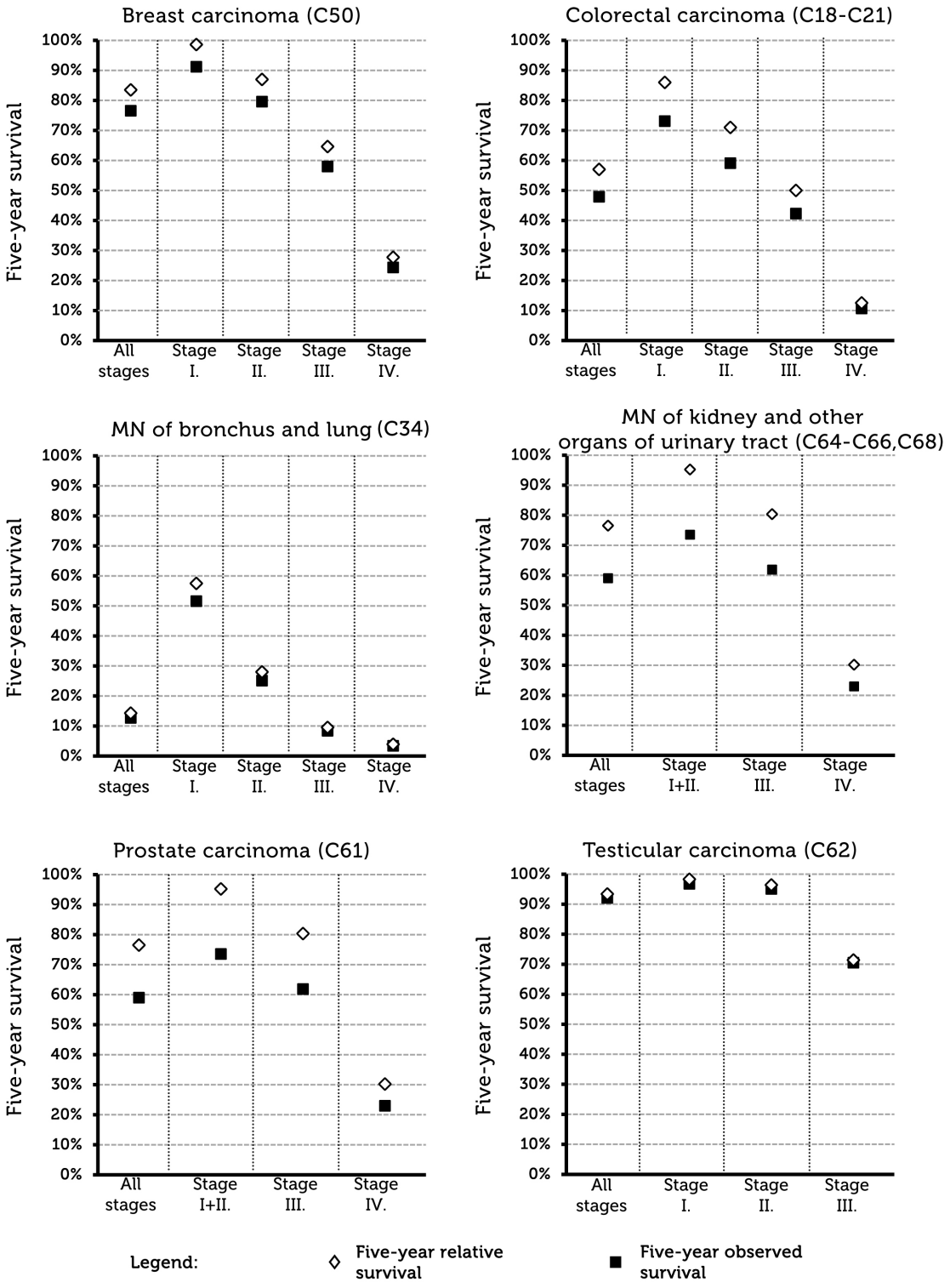

Figure 1.1: Five-year observed and relative survival rates in treated Czech cancer patients (selected diagnoses, 2003-2005 period analysis).

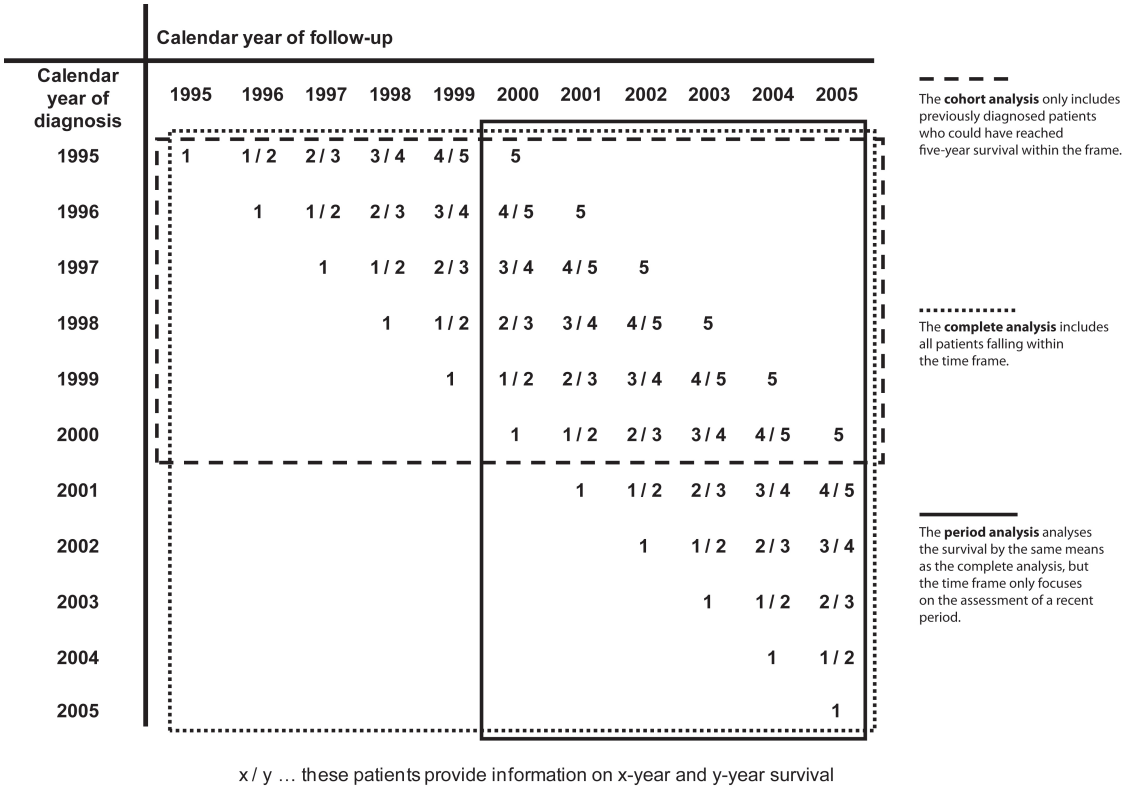

Figure 1.2: Methodical scheme for assessing the five-year survival rates, the period 1995-2005 being taken as an example.

(Figure 1.2). In this case, however, survival assessment for 2005 would be reported from the real situation in 1995–2000, which probably does not correspond to more recent results. For this reason, methods involving more recent data have been developed.

The complete analysis is an example of a method that takes into account more recent data [80]. Considering the example above, the complete analysis would involve all patients diagnosed in the period 1995–2005, regardless of the length of their follow-up. The survival estimate is derived from all patients recorded in the monitored period (Figure 1.2). The period analysis is another example [9]; this method also takes into account all patients diagnosed in the period 1995–2005, but the analysis only includes patients whose follow-up ended recently (the so-called "left-truncation"). This ensures that the survival estimate will be based on the most recent information on survival: patients diagnosed in the period 1999–2005 provide information on one-year survival rate, those diagnosed between 1998–2004 provide information on the two-year survival rate, etc. (Figure 1.2). A certain disadvantage of this method could be in overestimation of the survival rate, particularly in diagnoses where the actual trends in diagnostics cause also a

positive shift in proportional representation of the clinical stages [11].

1.2.5 Introduction of the next chapters: Outline of the thesis

The main aim of this dissertation is to present mathematical methods that can be used for modelling of cancer patient survival and demonstrate the usability of these methods on different real data sets, considering both population-based and individual cancer patient data.

The mathematical methods used for cancer survival assessment are introduced in Chapter 2. Some basic survival analysis methods are recalled and the concept of relative survival is presented. Then, Cox proportional hazards model is defined as the most popular model for cancer survival modelling together with the estimation principles of regression coefficients using the partial log likelihood. Special attention is given to methods for testing the proportional hazards assumption and methods of regression diagnostics appropriate for the Cox model. Furthermore, frailty models which can be considered as an extension of the Cox model are introduced as well as mixture cure fraction model for the modelling of relative survival.

In Chapter 3, a new model for the estimation of prevalence of patients requiring active anti-tumour therapy that is accessible from population-based cancer registry data is presented. The new model is an extension of a model already published in [27, 28] and [75]. Unlike the methods published so far, the new model 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 budget impact of the treatment even more influencing than age at diagnosis. The model is further applied on colorectal cancer data from the Czech National Cancer Registry to model the number of potentially treated patients with colorectal carcinoma in the Czech Republic in 2011.

Chapter 4 presents the survival benchmarks for Czech cancer patients with respect to stage at diagnosis and administration of anti-tumour treatment that has been published in [74]. Moreover, this chapter provides an overview of the observed and relative survival rates achieved since 1990. Moreover, a comparison is provided between the Czech and European data concerning the five-year relative survival rates in selected cancer diagnoses.

Finally, Chapter 5 presents a Cox regression model for the achievement of the complete cytogenetic or major molecular response to a modern targeted therapy in Czech and Slovak patients in chronic phase of chronic myeloid leukemia. The objective of this chapter is the identification of CML patient characteristics associated with prolonged time to complete cytogenetic response or major molecular

response to imatinib therapy, which could further indicate the increased risk of disease progression.

Mathematical methods for cancer survival assessment

2

This chapter presents mathematical methods that can be used for survival assessment of oncology patients. Some basic survival analysis methods are recalled and the concept of relative survival is introduced. Cox proportional hazards model is also addressed, and the estimation principles of regression coefficients using the partial log likelihood together with the methods for regression diagnostics are given. Furthermore, frailty models which can be considered as an extension of the Cox model are introduced as well as mixture cure fraction model for the modelling of relative survival.

2.1 The basics of survival analysis

In general, survival analysis refers to the collection of statistical methods used to study the time interval from a defined start of the follow-up to the moment at which the event of interest occurs [65]. When assessing cancer care, these events can vary in their definition; in population studies, the date of diagnosis is routinely considered as the starting point. The event of interest might be the patients death or, alternatively, disease recurrence or progression.

Let T be the positive real valued time variable with a continuous probability distribution and finite expectation. Considering T be the time to occurrence of some event in a population, we can define several interpretable functions that characterize the distribution of T:

- The probability density of T: $f(t), t \geq 0$.
- The survival function: $S(t) = \mathrm{P}(T > t) = \int_t^\infty f(x)dx = 1 - F(t)$, where $F(t)$ is the cumulative distribution function. Survival function describes the probability of surviving beyond a specified time t.
- The hazard function: $h(t) = f(t)/S(t) = \lim_{u \to 0} \frac{\mathrm{P}\{t<T\leq t+u\}/\mathrm{P}\{T>t\}}{u} = -\frac{d \ln S(t)}{dt}$. Hazard function at time t is defined as the ratio of the probability density function, $f(t)$, and the survival function, $S(t)$, and can be thought of as an instantaneous rate of a failure at time t.

- The cumulative hazard function: $H(t) = \int_0^t h(x)dx$. Cumulative hazard function refers to the "accumulation" of the hazard over time and can be used to estimate the survival function: $S(t) = \exp(-H(t))$.

Obviously, the event of interest does not have to occur in all individuals during the follow-up period. The time of survival is then referred to as censored, which means that the follow-up ended before the event of interest occurred. Censoring always brings about certain degree of information loss, [60], therefore, the proportion of censored patients becomes a certain qualitative measure of the input data. It is also important to note that censoring and the occurrence of the event of interest are assumed to be mutually independent, which means that the event of interest in censored patients is neither more nor less likely than in other patients. Thus, the censoring mechanism is assumed to be noninformative.

The methodology of population-based cancer survival analysis is principally similar as the methodology for survival analysis in other experimental areas. However, its purpose lies more in the description of patient survival in a demographically representative way [24]. This does not mean that we cannot make comparisons and inference in the population-based cancer survival analysis as a relevant prognostic factors such as clinical stage, localization, and histologic type are also available, but the primary goal of population-based analyses is the generalisation of results on the entire population of cancer patients. It follows immediately that such analyses require the analysed group to be representative enough of the general population.

2.1.1 Estimators of the survival function

Even if many parametric approaches have been formulated for the survival function estimation, the non-parametric methods are much more often applied in life and health sciences. Probably the best known non-parametric estimator of the survival function is the Kaplan-Meier (KM) estimator [58] that can be written as a product limit estimator of the following form:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R_i}\right). \tag{2.1}$$

where (t_i, d_i) represents data of ith individual, $i = 1, \ldots, n$. Time t_i is the observed follow-up time, i.e. either the time of failure or the censoring time, whereas d_i is the indicator variable with values $d_i = 1$, if failure has occurred, and $d_i = 0$, if censoring has been reported. R_i denotes the number of subjects at risk at time t_i, i.e. the number of subjects without failure and uncensored just before time t_i. KM estimator provides a point estimate which should be always accompanied

with confidence interval (CI) to describe the variability of this estimate. The variance of KM estimator needed for computation of CI can be estimated using the Greenwood's formula [43], which is given by

$$\widehat{\text{var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}. \tag{2.2}$$

In the population-based survival analysis, the survival rates are often calculated using life table method which is due to frequent inconsistencies in data quality and standard length of time interval for population-based survival analysis being one year. Let d_j denote the number of failures during the jth time interval, $j = 1, \ldots, J$, R_j be the number of individuals at risk at the start of interval j and c_j be the number of individuals whose survival time was censored during the jth time interval. Then the probability that a subject survives from beginning of the follow-up until the end of the Jth interval, known as the cumulative survival proportion and denoted here as $\hat{S}(J)$, is given by

$$\hat{S}(J) = \prod_{j=1}^{J} p(j) = \prod_{j=1}^{J} \left(1 - \frac{d_j}{R_j - c_j/2}\right). \tag{2.3}$$

Alternatively to the KM estimator, survival function can be derived using the cumulative hazard function which can be estimated in a non-parametric manner by the formula due to Nelson [68] and Aalen [1]:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{R_i}. \tag{2.4}$$

The standard estimate of variance for Nelson-Aalen estimator of $H(t)$ was originally proposed by Aalen [1] and is of the form:

$$\widehat{\text{var}}(\hat{H}(t)) = \sum_{t_i \leq t} \frac{d_i}{R_i^2}. \tag{2.5}$$

Let's consider data of the ith individual to be represented by (t_i, d_i), then because the observations are assumed to be independent, the likelihood function can be written as

$$L = \prod_{i=1}^{n} \mathrm{P}\{t_i, d_i\} = \prod_{i=1}^{n} S(t_i)^{1-d_i} f(t_i)^{d_i}. \tag{2.6}$$

2.2 Relative survival

Relative survival is calculated as the ratio of observed survival to the so-called expected survival rate (expressing mortality in the general population which corresponds to the monitored group of patients in terms of age and sex). The relative survival is a weighted equivalent of the observed survival, the weight being survival of the general population. The biggest advantage of relative survival is the fact that detailed records on causes of death in individual patients are not needed in order to calculate it [24].

Let $h^*(t)$ and $S^*(t)$ be the expected hazard function and the expected survival function of the general population, respectively, and similarly, let $h(t)$ and $S(t)$ be the observed hazard and survival function, respectively. Then the relative hazard function, denoted as $h^R(t)$, can be calculated as:

$$h^R(t) = h(t) - h^*(t), \tag{2.7}$$

and the relative survival function, denoted as $S^R(t)$, is given by:

$$S^R(t) = \frac{S(t)}{S^*(t)}. \tag{2.8}$$

It should be noted that the term relative survival proportion is not precise because in fact, it is not a proportion nor a rate but a ratio of two proportions. With respect to [23], the ratio can be defined as the result of dividing one quantity by another. On the other hand, a proportion can be defined as a type of ratio in which the numerator is included in the denominator, while a rate is a measure of change in one quantity per unit of another quantity on which the first quantity depends. The relative survival proportion, $S^R(t) = S(t)/S^*(t)$, is thus a ratio of two proportions.

2.2.1 Methods for estimation of expected survival

The principle of expected survival estimation is the calculation of survival proportion for a comparable group from the general population that can be regarded as practically free of the disease of interest. The expected survival proportion estimates are based on population life-tables or, more specifically, on annual probabilities of death in the general population, that are matched to the subjects under study according to age, sex, and calendar time. The calculation of the expected survival proportion for a group of cancer patients involves calculating the expected survival probability for each individual cancer patient. Two methods are commonly utilised for the estimation of expected survival, both having different advantages and purpose of use.

I. The Ederer II method

Let $S^*(J)$ be the cumulative expected survival proportion of the whole group up to the end of the Jth interval, and $S_i^*(J)$ be the same for the ith individual. Similarly, define $p^*(j)$ as the interval-specific expected survival proportion of the group for the jth interval and $p_i^*(j)$ as the same quantity for the ith individual. Then $S^*(J)$ can be calculated according to Ederer and Heise [31] by

$$S^*(J) = \prod_{j=1}^{J} p^*(j), \tag{2.9}$$

where

$$p^*(j) = \sum_{i=1}^{R_j} p_i^*(j)/R_j \tag{2.10}$$

is the average of the annual expected survival probabilities $p_i^*(j)$ of the patients at risk at the start of the jth interval. The Ederer II estimates of the interval-specific expected survival proportions are thus based on only those patients at risk at the start of the interval, that means, this method allows for heterogeneous observed follow-up times. However, Hakulinen has shown in [45] that the cumulative expected survival proportion is then dependent on the observed mortality, which leads to biased estimates (usually underestimates) of the relative survival ratio.

II. The Hakulinen method

An alternative method for the expected survival estimation was proposed by Hakulinen [45] which has become a standard for the estimation of cumulative expected survival for the purpose of estimating relative survival ratios in population analysis till then. This method adjusts the estimates for potentially heterogeneous follow-up times among the subjects making the estimates independent of the observed mortality of the patients. The adjustment is performed through the use of the so-called potential follow-up time that should be specified for each individual. Fact that the potential follow-up times are required for all individuals can be problematic within studies where individuals have different last days of contact, especially for deceased patients. Fortunately, this is not a problem in studies evaluating population-based cancer registry data for the population-based registries usually have a common closing date.

The calculation of the Hakulinen expected survival estimates is performed as follows. Let k_j denote the number of patients with a potential follow-up time extending beyond the start of the jth time interval. Let $k_j = k_{j,a} + k_{j,b}$, where $k_{j,a}$ is the number of patients with a potential follow-up time extending beyond

the end of the jth time interval and $k_{j,b}$ is the number of potential withdrawals during the jth time interval. It follows that $k_1 = R_1$, and $k_{j+1} = k_{j,a}$. With K_j, $K_{j,a}$, and $K_{j,b}$ representing the set of k_j, $k_{j,a}$, and $k_{j,b}$ patients, respectively, then the expected number of patients alive and under the risk at the beginning of the jth time interval can be written as:

$$R_j^* = \begin{cases} \sum_{i \in K_j} S_i^*(j-1) & \text{for } j \geq 2 \\ R_1 & \text{for } j = 1 \end{cases} \tag{2.11}$$

We further assume that each of the $k_{j,b}$ patients with the follow-up time ending during the jth time interval is at risk for the half of the interval, implying that the expected probability of dying in the jth time interval is $1 - \sqrt{p_i^*(j)}$. Thus, the number of patients censored alive in the jth time interval is given by:

$$c_j^* = \begin{cases} \sum_{i \in K_{j,b}} S_i^*(j-1)\sqrt{p_i^*(j)} & \text{for } j \geq 2 \\ \sum_{i \in K_{1,b}} \sqrt{p_i^*(1)} & \text{for } j = 1 \end{cases} \tag{2.12}$$

The rest of the $k_{j,b}$ patients are assumed to die during the jth interval, so the expected number of patients dying in the jth time interval, among potential withdrawals during the jth interval, is estimated as:

$$\delta_j^* = \begin{cases} \sum_{i \in K_{j,b}} S_i^*(j-1)\left[1 - \sqrt{p_i^*(j)}\right] & \text{for } j \geq 2 \\ \sum_{i \in K_{1,b}} \left[1 - \sqrt{p_i^*(1)}\right] & \text{for } j = 1 \end{cases} \tag{2.13}$$

The δ_j^* quantity is then added to the expected number of patients dying among the $k_{j,a}$ patients, resulting in the expected total number of patients dying in the jth interval given by:

$$d_j^* = \begin{cases} \{\sum_{i \in K_{j,a}} S_i^*(j-1)[1 - p_i^*(j)]\} + \delta_j^* & \text{for } j \geq 2 \\ \{\sum_{i \in K_{1,a}} [1 - p_i^*(1)]\} + \delta_1^* & \text{for } j = 1 \end{cases} \tag{2.14}$$

Then the interval-specific expected survival proportion of the group in the jth interval is estimated by:

$$\hat{p}^*(j) = 1 - \frac{d_j^*}{R_j^* - c_j^*/2}, \tag{2.15}$$

and the cumulative expected survival proportion of the whole group from the follow-up start up to the end of the Jth interval is given by

$$\hat{S}^*(J) = \prod_{j=1}^{J} \hat{p}^*(j). \tag{2.16}$$

It should be noted that the principal assumption of the Hakulinen's method is that information on potential follow-up times are available for all patients. Thus, the Hakulinen method can be smoothly used in population-based cancer registry data analyses for the potential follow-up time can be easily calculated based on one follow-up closing date common for all patients. However, the use of Hakulinen's method may be limited in clinical or cohort studies when there are individual dates of last contact.

2.2.2 Confidence intervals for relative survival

The relative survival estimates should be always accompanied with confidence intervals as it is a standard method for presenting the amount of random error in an estimate. As the variance of the expected survival proportion can be assumed negligible compared to the variance of the observed survival proportion, the expected survival proportion is assumed to be a constant value [23]. Then the variance of the relative survival ratio (both interval-specific and cumulative) can be expressed as follows

$$\text{var}\left(S^R(t)\right) = \text{var}\left(\frac{S(t)}{S^*(t)}\right) = \frac{\text{var}(S(t))}{S^*(t)^2} = \frac{SE(S(t))^2}{S^*(t)^2}. \tag{2.17}$$

A 95% CI can be then constructed for the relative survival estimate, $\hat{S}^R(t)$, as $\hat{S}^R(t) \pm u_{0.975} \times \hat{SE}(\hat{S}^R(t))$ with $\hat{SE}(\hat{S}(t))$ in (2.17) estimated using the Greenwood's formula [43] given by (2.2), and $u_{0.975}$ being the 0.975 quantile of the standard normal distribution.

2.2.3 Age standardisation

The calculation of relative survival itself does not guarantee that these estimates would be comparable among the different populations of cancer patients, particularly if they differ in their age structures. In this case, age-specific relative survival must be calculated; this is the calculation of relative survival in several age categories and a subsequent weighting of these partial estimates with the weights of corresponding age groups [10]. When describing a given population, weights corresponding to individual age groups are most frequently specified with respect to the relative proportion of that particular age group in standard population of cancer patients. Given that W_k is the weight of the kth age group and R_k is the age-specific relative survival of that age group, the age-standardized relative survival (ASRS) is calculated as a weighted average according to the following formula [19]:

$$ASRS = \frac{\sum_k W_k R_k}{\sum_k W_k}. \tag{2.18}$$

It is obvious that the definition of age groups and corresponding weights is a key step which might significantly affect the results of the calculation. In comparative analyses, equal weights must be used for age standardisation of groups under comparison [6, 18, 19].

2.3 Cox proportional hazards model

The Cox proportional hazards model is the most popular model for the analysis of time to event data in medicine as it allows the survival probability to depend not only on time, but also on the vector of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$, as in common regression analysis while having nice interpretational properties. The hazard function for patient indexed with i can be written as follows

$$h(t, \mathbf{x}_i) = h_0(t) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p), \tag{2.19}$$

where $h_0(t)$ is the so-called baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of regression coefficients corresponding to the vector of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$. The model can be seen as a modification of a parametric model based on exponential distribution, however, unlike the exponential model the Cox model leaves the baseline hazard function $h_0(t)$ unspecified, and thus it is not a fully parametric model. In this setting, it can be seen that the hazards for two groups of patients are proportional in a following way

$$\frac{h(t, \mathbf{x}_i)}{h(t, \mathbf{x}_j)} = \frac{h_0(t) \exp(\mathbf{x}_i'\boldsymbol{\beta})}{h_0(t) \exp(\mathbf{x}_j'\boldsymbol{\beta})} = \exp((\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}), \tag{2.20}$$

i.e. the proportion of hazards of two individuals is constant in time. This is both the basic assumption of the Cox model as well as the principle for presentation of results and their interpretation. The most frequently used output of Cox modelling are the estimates of excess hazards associated with variables of interest.

2.3.1 Estimation of the regression parameters

Derivation of an estimator of $\boldsymbol{\beta}$ cannot be based on an ordinary likelihood function since $h_0(t)$ is not specified parametrically in the Cox model. Instead, the so-called *partial* likelihood has been proposed by Cox [20] for the estimation of regression parameters which is a function depending on $\boldsymbol{\beta}$ only. Consider a sample of n subjects with a total of k failures ($k \leq n$). Furthermore, let $t_1 < t_2 < \ldots < t_k$

be the k distinct ordered failure times observed and R_i be the set of individuals at risk of failing just before failure time t_i. Considering that the vector of covariates $\mathbf{x}_i$ of the ith individual is constant in time, the conditional probability that individual i is observed to fail at t_i, given that only one failure occurs at t_i, is

$$\frac{h(t_i, \mathbf{x}_i)}{\sum_{j \in R_i} h(t_i, \mathbf{x}_j)} = \frac{h_0(t_i) \exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{j \in R_i} h_0(t_i) \exp(\mathbf{x}_j' \boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})}. \tag{2.21}$$

Assuming that these conditional probabilities are conditionally independent across the different failure times, the partial likelihood function covering the failure pattern of the whole set of n subjects with a total of k failures can be computed as the product over k observed failure times:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})}. \tag{2.22}$$

The regression coefficients $\boldsymbol{\beta}$ are estimated with $\hat{\boldsymbol{\beta}}$ that maximize the partial likelihood, $L(\boldsymbol{\beta})$, or its logarithm, $\log L(\boldsymbol{\beta})$, where notation $\log$ stands for the natural logarithm (as well as in the rest of the thesis). The partial log likelihood takes the following form [65]

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{k} \left\{ \mathbf{x}_i' \boldsymbol{\beta} - \log \left[\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta}) \right] \right\} = \sum_{i=1}^{k} l_i. \tag{2.23}$$

The estimates of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ are then derived by equating the p first derivatives of $\log L(\boldsymbol{\beta})$ to zero (with respect to $\beta_m, m = 1, \ldots, p$) and solving this system of equations with an iterative method, e.g. the Newton-Raphson algorithm. The derivatives of the contribution l_i will be also further utilised for the formulation of tests about the vector of regression coefficients, $\boldsymbol{\beta}$. Let $\mathbf{x}_i$ be the vector of covariates of the subject experiencing the failure at time t_i, and x_{im} be its mth component. Then the first derivative of the contribution l_i with respect to β_m is

$$\frac{\partial l_i}{\partial \beta_m} = x_{im} - \frac{\sum_{j \in R_i} x_{jm} \exp(\mathbf{x}_j' \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})}, \tag{2.24}$$

whereas the second derivative of l_i with respect to β_m can be written as

$$\frac{\partial^2 l_i}{\partial \beta_m^2} = \left[- \frac{\sum_{j \in R_i} x_{jm}^2 \exp(\mathbf{x}_j' \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})} - \left(\frac{\sum_{j \in R_i} x_{jm} \exp(\mathbf{x}_j' \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})} \right)^2 \right], \tag{2.25}$$

The first derivative is the difference between the value of the mth covariate of the individual failing at t_i, and the weighted average of the mth covariate taken over

the subjects at risk at t_i, namely over the set R_i. The sum of the first derivatives over all failure times is usually denoted as $U_m(\boldsymbol{\beta})$, i.e. $U_m(\boldsymbol{\beta}) = \sum_{i=1}^{k} \partial l_i / \partial \beta_m$, and together with the remaining $m-1$ components form the score vector $\mathbf{U}(\boldsymbol{\beta})$, i.e. $\mathbf{U}(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), U_2(\boldsymbol{\beta}), \ldots, U_m(\boldsymbol{\beta}), \ldots, U_p(\boldsymbol{\beta}))'$. It is evident that $\mathrm{E}(U_m) = 0$. The second derivative of the log likelihood contribution given by equation (2.25) has the form of variance and its negative value is the (im)th element of the observed information matrix, so

$$\mathbf{I}(\boldsymbol{\beta}) = \begin{vmatrix} \frac{\partial l_1}{\partial \beta_1} & \cdots & \frac{\partial l_1}{\partial \beta_p} \\ \vdots & \frac{\partial l_i}{\partial \beta_m} & \vdots \\ \frac{\partial l_k}{\partial \beta_1} & \cdots & \frac{\partial l_k}{\partial \beta_p} \end{vmatrix} \tag{2.26}$$

The inverse of the information matrix, evaluated at $\hat{\boldsymbol{\beta}}$, can be used as the estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, i.e.

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}). \tag{2.27}$$

In general, the interpretation of a Cox model involves examining the regression coefficients for each individual variable. A positive regression coefficient for an variable means that the hazard is higher, and thus the survival prognosis worse, for higher values of that variable. On the contrary, a negative regression coefficient implies a better prognosis for individuals with higher values of that variable. Moreover, the interpretation of the hazard ratio depends on the character of considered clinical variable. In case of dichotomous variable, the interpretation of resulting excess hazard is very simple for it is readily attributable to the risk category. On the other hand, when considering continuous variable, the resulting value of hazard rate stands for excess hazard attributable to one unit increase in that variable.

Having the estimate of $\boldsymbol{\beta}$, we also need an estimate of the baseline hazard function to be able to obtain fitted hazards and survival functions for any value of $\mathbf{x}$. An estimator for the baseline hazard function, $H_0(t)$, which is highly referred to and used, was originally proposed by Breslow [13], and is given by

$$\hat{H}_0(t) = \sum_{t_i \le t} \hat{h}_0(t_i) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}})}, \tag{2.28}$$

where d_i is the number of failures at t_i which can be greater than 1 if there are ties in the observed failure times.

2.3.2 Handling tied failure times

The regression coefficients derivation process in the Cox model implicitly assumes that exact failure or censoring time is known for each subject, i.e. none two observed failure times are the same. However, the real data sets often contain tied failures that introduce complications to partial likelihood computation. Two computationally feasible approximation algorithms were previously proposed for partial likelihood derivation in case of tied failure times:

- **Breslow's approximation**. When the number of tied failure times is not large, Breslow's approximation [53, 90] of the log likelihood function can be used. It is given by $\log L(\boldsymbol{\beta}) = \sum_{i=1}^{k} \{\mathbf{s}_i' \boldsymbol{\beta} - d_i \log[\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})]$, where $\mathbf{s}_i = \sum_{j \in D_i} \mathbf{x}_j$, D_i is the set of individuals failing at time t_i, and d_i is the number of failures at t_i.

- **Efron's approximation**. Efron derived another approximation of the partial log likelihood which provides a better approximation to the exact likelihood than the Breslow's approximation [53]. It is given by $\log L(\boldsymbol{\beta}) = \sum_{i=1}^{k} \{\mathbf{s}_i' \boldsymbol{\beta} - \sum_{l=1}^{d_i} \log[\sum_{j \in R_i} \exp(\mathbf{x}_j' \boldsymbol{\beta}) - \frac{l-1}{d_i} \sum_{j \in D_i} \exp(\mathbf{x}_j' \boldsymbol{\beta})]$.

The exact partial likelihood can also be derived for failure time data with ties. The idea of its computation is that for each failure time t_i the number of failures d_i can be ordered in $(d_i)!$ different ways, with the average value over this set of permutations taken as the final value of the partial likelihood. It follows that if the number of ties is large computation of the exact partial likelihood becomes computationally intensive. More details can be found in [56].

2.3.3 Wald, score and likelihood ratio tests

There are three asymptotic tests standardly defined for the statistical inference on $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$: the partial likelihood ratio test, the Wald test and the score test [90]. Let us say we want to test the hypothesis that r components of the $\boldsymbol{\beta}$ vector are null, in which case they could be dropped out of the model. Obviously, we can assume that the r coefficients to be tested are the first r coefficients in the vector $\boldsymbol{\beta}$ without loss of generality. Then the hypothesis to be tested is

$$H_0 = \beta_1 = 0, \ldots, \beta_r = 0. \tag{2.29}$$

Then the vector $\boldsymbol{\beta}$ can be rearranged as follows: $\boldsymbol{\beta} = (\boldsymbol{\beta}^*, \boldsymbol{\beta}^{**})$. The mentioned tests are all closely related to the partial likelihood function [65] and can be defined as follows:

- **The likelihood ratio test**: this test can be defined as twice the difference in the log partial likelihood at the actual estimate of $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}^{**})$ corresponding to the observed data and at the vector $(\mathbf{0}, \tilde{\boldsymbol{\beta}}^{**})$ corresponding to the null hypothesis. The test statistic can be then written as

$$Q_{LR} = 2(\log L(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}^{**}) - \log L(\mathbf{0}, \tilde{\boldsymbol{\beta}}^{**})), \tag{2.30}$$

  and is asymptotically distributed as χ^2 with r degrees of freedom.

- **The Wald test**: this test is based on the maximum likelihood estimate of $\boldsymbol{\beta}^*$ under the full model. The test statistic of the Wald test is given by

$$Q_W = (\hat{\boldsymbol{\beta}}^* - \mathbf{0})'(\mathbf{I}^{-1}_{r\times r}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}^{**}))^{-1}(\hat{\boldsymbol{\beta}}^* - \mathbf{0}), \tag{2.31}$$

  where $\mathbf{I}^{-1}_{r\times r}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}^{**})$ is a submatrix of dimension $r \times r$ of the entire variance matrix corresponding to $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}^{**})$ estimated under the full model. The test statistic of the Wald test is also distributed as χ^2 with r degrees of freedom.

- **Rao's score test**: this test is based on the log likelihood evaluated at the maximum likelihood estimate of $\boldsymbol{\beta}^{**}$ under the restricted model. More specifically, it is based on the gradient of the log likelihood function at $\boldsymbol{\beta}^* = \mathbf{0}$, i.e. on the $r \times r$ score vector:

$$\mathbf{U}_{H_0} = \mathbf{U}(\mathbf{0}, \tilde{\boldsymbol{\beta}}^{**}) = \left[\frac{\partial \log L(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{**})}{\partial \beta_1}, \ldots, \frac{\partial \log L(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{**})}{\partial \beta_r}\right]'_{\boldsymbol{\beta}^*=\mathbf{0}, \boldsymbol{\beta}^{**}=\tilde{\boldsymbol{\beta}}^{**}} \tag{2.32}$$

  evaluated at $\boldsymbol{\beta}^* = \mathbf{0}$ and $\boldsymbol{\beta}^{**} = \tilde{\boldsymbol{\beta}}^{**}$. The test statistic of the Rao's score test can be written as

$$Q_R = \mathbf{U}'_{H_0}\mathbf{I}^{-1}_{r\times r}(\mathbf{0}, \tilde{\boldsymbol{\beta}}^{**})\mathbf{U}_{H_0}, \tag{2.33}$$

  where $\mathbf{I}^{-1}_{r\times r}(\mathbf{0}, \tilde{\boldsymbol{\beta}}^{**})$ is a submatrix of dimension $r \times r$ of the inverse of the observed information matrix evaluated at $(\mathbf{0}, \tilde{\boldsymbol{\beta}}^{**})$. Under the null hypothesis, the Rao's score test statistic follows a χ^2 distribution with r degrees of freedom.

2.3.4 Assessment of the proportional hazards assumption

Proportionality of hazards associated with individual model variables is the crucial assumption of the Cox model. This necessary condition is often appropriate for survival data but every time the model is used this assumption should be

verified. As already mentioned, the effect of the predictors in the Cox model is assumed to be the same at all times t, i.e. the log hazard function is of the form

$$\log h(t, \mathbf{x}) = \log h_0(t) + \mathbf{x}'\boldsymbol{\beta}. \tag{2.34}$$

As an example, let's consider that the model include one dichotomous covariate, coded as 0 or 1, respectively. Then a plot of the log hazard function would contain two curves, $\log h_0(t)$ for $x = 0$ and $\log h_0(t) + \beta$ for $x = 1$, i.e. regardless of how complicated the baseline function, $h_0(t)$, is, the difference between these two curves is β for any t. As for continuous covariates, consider age as a single variable in the model and assume we are interested in two log hazard functions for age a and $a + 10$. Then, if the coefficient, β, is positive, the difference between these two curves is 10β for any t.

The rationale behind the verification of proportional hazards assumption is to assess the extent to which the log hazard functions are equidistant from each other in time. Various methods have been proposed for testing the proportionality assumption, three of them commonly used in survival analysis will be described here in more detail:

- Graphical check of the proportional hazards
- Test based on time-dependent covariates
- Test based on scaled Schoenfeld residuals

1. Graphical check of the proportional hazards

A simple graphical check of the proportional hazards can be performed using the above mentioned principle of two equidistant log hazard functions that can be equivalently expressed using the following consideration: The survival function with respect to cth covariate satisfies under the Cox model

$$S_c(t) = \exp(-H_0(t)\exp(x_c\beta_c)),\ c = 1, \ldots, p, \tag{2.35}$$

and therefore also

$$\log[-\log(S_c(t))] = \log H_0(t) + x_c\beta_c, \tag{2.36}$$

so if the assumption is correct the log cumulative hazards, $\log[-\log(S_c(t))]$, for the levels of covariate c should appear to be approximately parallel.

2. Test based on time-dependent variables

Cox in his original paper proposed also a simple test for proportional hazards assumption based on defined time-dependent covariates [20]. Let's consider a dichotomous covariate x_c to be tested for proportional hazards and the vector of $p-1$ remaining covariates $\mathbf{x}_{-c}$. Then we can define a time-dependent covariate $x_c(t) = x_c g(t)$ which is clearly a transformation of x_c. The time-dependent function, $g(t)$, is usually chosen as the identity function, $g(t) = t$, or the logarithmic function, $g(t) = \log t$.

This notation leads to the hazard ratio of two subjects with values $x_c = 1$ and $x_c = 0$ and the same covariates $\mathbf{x}_{-c}$ of the form $\exp(x_c\beta_c + x_c(t)\gamma)$ instead of $\exp(x_c\beta_c)$ assumed in (2.20). The test on the proportional hazards is then equivalent with testing for a non-zero value of γ which would imply that the hazard ratio associated with x_c changes in time.

3. Test based on scaled Schoenfeld residuals

In 1994 Grambsch & Therneau [40] proposed a test on proportional hazards based on scaling of residuals first introduced by Schoenfeld [81]. Let $t_1 < t_2 < \ldots < t_k$ be the k distinct ordered failure times, and let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ and $R_1, \ldots, R_k$ be the corresponding vector of covariates and risk set. The Schoenfeld's residuals can then be defined as:

$$\hat{\mathbf{r}}_i^S = \mathbf{x}_i - \frac{\sum_{j \in R_i} \mathbf{x}_j \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}})}{\sum_{j \in R_i} \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}})} = \mathbf{x}_i - \hat{\bar{\mathbf{x}}}_{w_i}, \tag{2.37}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$, derived by maximising of (2.23). Let the approximate estimate of the variance matrix of the vector of Schoenfeld residuals be denoted as $\widehat{\text{var}}(\hat{\mathbf{r}}_i^S)$. Then, for the ith uncensored individual (the estimator is missing for censored subjects), the elements in this matrix are

$$\widehat{\text{var}}(\hat{\mathbf{r}}_i^S)_{lm} = \hat{v}_{lm}^i = \sum_{j \in R(t_i)} \hat{w}_{ij}(x_{jl} - \hat{\bar{x}}_{w_i l})(x_{jm} - \hat{\bar{x}}_{w_i m}), \tag{2.38}$$

where $\hat{w}_{ij} = \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}}) / \sum_{m \in R(t_i)} \mathbf{x}_m' \hat{\boldsymbol{\beta}}$, and $\hat{\bar{x}}_{w_i l}$ is the lth element corresponding to the lth covariate of vector $\hat{\bar{\mathbf{x}}}_{w_i}$ defined in (2.37). Having the $\hat{\mathbf{r}}_i^S$ variance estimator, the scaled Schoenfeld residuals can be defined as

$$\hat{\mathbf{r}}_i^{S*} = \left[\widehat{\text{var}}(\hat{\mathbf{r}}_i^S)\right]^{-1} \hat{\mathbf{r}}_i^S. \tag{2.39}$$

Grambsch & Therneau considered an alternative to Cox proportional hazards model, a model with time-varying coefficients, $\beta_l(t) = \beta_l + g_l(t)\gamma_l$, where $g_l(t)$

is an unknown function of time which vary about zero, and γ_l is a regression coefficient. Next, they showed that the mean of the scaled Schoenfeld residuals at time t_i can be, for the lth covariate, approximated as follows:

$$\mathrm{E}(\hat{r}_{il}^{S*}) \approx g_l(t_i)\gamma_l, \tag{2.40}$$

Therefore, the function $g_l(t)$, for covariate l, can be estimated by a smoothed plot of the lth component of $\hat{\mathbf{r}}_i^{S*}$ against t_i. A formal test was also proposed by Grambsch & Therneau [90], considering a standard linear model for the $\hat{r}_{il}^{S*}$ values. The test statistic for the lth covariate can be written as

$$T(g_l) = \frac{(\sum_{i=1}^{k}\{g_j(t_i) - \bar{g_j}(t)\}\hat{r}_{il}^{S})^2}{\sum_{i=1}^{k}\hat{v}_{ll}^{i}(g_j(t_i) - \bar{g_j}(t))^2 - (\sum_{i=1}^{k}\{g_j(t_i) - \bar{g_j}(t)\}\hat{v}_{ll}^{i})^2 / \sum_{i=1}^{k}\hat{v}_{ll}^{i}}, \tag{2.41}$$

where n is the overall number of individuals, k is the total number of failures, d_i is the indicator variable of failure, and $\bar{g_j}(t)$ is the mean of the $g_j(t_i)$s. This test statistic is asymptotically distributed as a χ^2 with 1 degree of freedom. More details can be found in [40].

2.3.5 Stratification

Stratification can be regarded as an extension of model given by (2.19) that allows to adjust for a factor showing nonproportional hazards without estimating its effect [65]. Let us consider a factor with J levels, then the stratified Cox model according to this factor is given by

$$h_j(t, \mathbf{x}) = h_{0j}(t)\exp(\mathbf{x}'\boldsymbol{\beta}), \tag{2.42}$$

where j denotes the particular stratum ($j = 1, \ldots, J$). Under the stratified model, it can be seen that individuals within the jth stratum share the same baseline hazard function, $h_{0j}(t)$, which implies that the proportional hazards for two individuals in the same stratum still holds:

$$\frac{h_j(t, \mathbf{x}_1)}{h_j(t, \mathbf{x}_2)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\beta}). \tag{2.43}$$

On the other hand, individuals from different groups can have nonproportional hazards as their baseline hazards functions may differ. Computationally, the stratified Cox model is a generalisation of (2.23) in a way that the overall log likelihood becomes a sum of J log likelihoods incident to individual strata, i.e. the overall log likelihood is given by

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{J} \log L_j(\boldsymbol{\beta}). \tag{2.44}$$

2.3.6 Assessment of a model fit

Model validation is an important step in the model building process, and is usually performed using model residuals that represent the differences between the responses observed at each combination of the explanatory variables and the corresponding prediction of the response computed using the estimated regression function. However, the residuals in Cox regression model are not as useful for global model fit assessment as are residuals in linear models or another types of parametric models [90]. On the other hand, the residuals defined for Cox model can be used for specific purposes, e.g. for identification of poorly predicted individuals or influential observations.

There are three highly used residuals besides the Schoenfeld's defined above in the Cox model: the martingale, deviance and score residuals. Martingale residuals can be used for overall test of the goodness-of-fit of a Cox model [73] whereas deviance residuals, which can be derived from martingale residuals, can be used for detection of poorly predicted individuals. Score residuals can be, on the other hand, useful for determination of influential observations.

Considering only time-fixed covariates for simplicity, a martingale residual is defined for each individual as

$$\hat{r}_i^M = N_i - \hat{E}_i = N_i - \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{H}_0(\hat{\boldsymbol{\beta}}, t_i) \tag{2.45}$$

where N_i is the number of failures of the ith individual at time t_i (N_i is either 0 or 1 for the single-event survival data), $\hat{E}_i$ stands for the expected number of failures based on the estimated vector of regression coefficients, $\hat{\boldsymbol{\beta}}$, and $\hat{H}_0(\hat{\boldsymbol{\beta}}, t_i)$ is the Breslow's estimator of the baseline cumulative hazard function [13]. It can be seen that the martingale residual represents for each individual the observed number of events minus the number predicted by the fitted model and given the follow-up time t_i, i.e. it measures the contrast between the prediction and the reality.

To perform an overall test of the goodness-of-fit of a Cox model, Parzen & Lipsitz [73] considered a partition of the subjects into K groups according to their risk of failure expressed by $\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$, and proposed an alternative Cox model using this kind of information. Furthermore, they proposed the score statistic the goodness-of-fit, and showed that this statistic is actually a function of the martingale residuals within each group. They conclude that for *sufficiently large* sample sizes (the criteria for *sufficiently large* sample sizes is similar to those of Pear-

son's chi-square statistic for contingency tables) the statistic has approximately chi-square distribution with $K-1$ degrees of freedom.

As the martingale residuals are skewed, some authors prefer to use other residuals for the lack of fit assessment [60]. The deviance residuals, $\hat{r}_i^D$, which can be considered as a normalizing transformation of the martingale residuals [90], can be used this way, for example. It can be shown with a one-term Taylor expansion that

$$\hat{r}_i^D = \frac{N_i - \hat{E}_i}{\sqrt{\hat{E}_i}}. \tag{2.46}$$

A good fit of the Cox model can be anticipated when the plotted deviance residuals are scattered around zero. Residuals which are far from zero belong to individuals with a poor prediction. However, there is no definite value to be used as a threshold for specifying poorly predicted individuals as the deviance residuals have no reference probability distribution [90].

Another residuals that can be defined using the martingale residuals are the score residuals [53] expressing each individual's contribution to the score vector, $\mathbf{U}(\boldsymbol{\beta})$. The score residuals of the ith individual can be expressed as

$$\mathbf{r}_i^C = \sum_{j=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}_{w_j}\right) dr_i^M(t_j) \tag{2.47}$$

where $\bar{\mathbf{x}}_{w_j}$ is the average covariate vector over individuals at risk in time t_j defined in equation (2.37). The expression $dr_i^M(t_j)$ stands for the change in the martingale residuals for the ith subject at time t_j, and can be expressed as

$$dr_i^M(t_j) = dN_i(t_j) - Y_i(t_j)h_0(t_j)\exp(\mathbf{x}_i'\boldsymbol{\beta}), \tag{2.48}$$

where $dN_i(t_j)$ is the change in the count function of the ith individual at time t_j. It is always zero for censored individuals, whereas for uncensored individuals, it is equal to zero except at the observed time of failure, when $dN_i(t_i) = 1$. The function $Y_i(t_j)$ represents the *at risk process* of the ith individual and is given by

$$Y_i(t_j) = \begin{cases} 1 \text{ if } t_i \geq t_j \\ 0 \text{ if } t_i < t_j \end{cases}. \tag{2.49}$$

Finally, the function $h_0(t_j)$ represents an increment of the Breslow's estimator of $H_0(t)$ evaluated at t_j. The estimate of score residuals can be written with respect to (2.48) as

$$\hat{\mathbf{r}}_i^C = d_i(\mathbf{x}_i - \hat{\bar{\mathbf{x}}}_{w_j}) - \sum_{j=1}^{n}(\mathbf{x}_i - \hat{\bar{\mathbf{x}}}_{w_j})Y_i(t_j)\hat{h}_0(t_j)\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}). \tag{2.50}$$

Obviously, the score residuals of ith individual form a row vector of length p with components $\hat{r}_{ij}^C, j = 1, \ldots, p$, that is $\hat{\mathbf{r}}_i^C = \left(\hat{r}_{i1}^C, \hat{r}_{i2}^C, \ldots, \hat{r}_{ip}^C\right)'$. However, a scaled version of the score residuals is used as a measure of influence of an individual observation on a maximum partial likelihood estimate of a regression coefficient. The scaled score residuals are defined as

$$\hat{\mathbf{r}}_i^{C*} = \widehat{\text{var}}(\hat{\boldsymbol{\beta}})\hat{\mathbf{r}}_i^C, \tag{2.51}$$

where $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ is the estimate of the variance matrix of the regression coefficient estimates introduced in (2.27).

It should be noted that the martingale and the deviance residuals are subject-specific by focusing at precision of model prediction for each individual. On the other hand, score and Schoenfeld residuals are rather covariate-specific as they primarily focus on the difference between covariate vectors.

2.3.7 Competing risks in Cox regression

In many clinical experiments, with cancer studies being no exception, failures can be divided into two or more groups according to several distinct causes. These different causes of failure should be considered as competing events, where occurrence of any one of the events causes failure and precludes the occurrence of the other events. Such situation introduces the so-called competing risks that can be represented with the so-called cause-specific hazard functions. Consider C be the observed cause of failure, and let us say that there are J different competing causes of failure, then the hazard function specific for the jth cause of failure can be defined as

$$h_j(t) = \lim_{u \to 0} \frac{\text{P}\{t < T \leq t + u, C = j\}/\text{P}\{T > t\}}{u}, \text{ for } j = 1, \ldots, J. \tag{2.52}$$

The $h_j(t)$ function can be interpreted as the instantaneous failure rate of cause j at time t. As a rate, the value of a cause specific hazard function has the unit of probability per time unit. A Cox proportional hazards model can be considered for the cause-specific hazard function, i.e., given the $p \times 1$ vector of covariates $\mathbf{x}$, the hazard function specific for the jth cause can be modelled as $h_j(t, \mathbf{x}) = h_{0j}(t) \exp(\mathbf{x}'\boldsymbol{\beta})$.

If one is interested only in one failure cause and the censoring is independent, then the methodology of partial likelihood can be used for analysing the competing risks data, fitting a Cox model for the failure type of interest and treating the other causes of failure as censored observations [56]. However, if we are interested in the comparison of parameter estimates corresponding to different

causes of failure, and in estimation of the ratio between pairs of baseline hazard functions, an alternative approach need to be adopted.

Lunn & McNeil [63] published a method for fitting Cox regression model that can cope with competing risks by augmenting the data using a duplication method. The idea behind their method is that the hazard functions of different failure types are assumed to be additive, and thus the overall hazard function for all failure types is the sum of the particular risk processes. When failure of a specific type is reported, the observed failure time corresponds to the minimum of the failure times associated with these processes, with the particular risk process being uncensored, and the rest of them being censored. Their approach can be best demonstrated on an example. Suppose, for simplicity, that there are only two types of failure, say I and II, denoted with an indicator, say c, which means that $c = 0$ for type I, and $c = 1$ for type II, respectively. Then an individual i with failure time t_i, vector of covariates $\mathbf{x}_i$, and failure type c_i should be represented in data with the following entries:

ID	Time	Failure status	Failure type	Covariates
i	t_i	1	c_i	$\mathbf{x}_i$, $c_i\mathbf{x}_i$
i (replicate)	t_i	0	$1 - c_i$	$\mathbf{x}_i$, $(1 - c_i)\mathbf{x}_i$

The Cox model is then applied on the duplicated covariates, with failure type, c, being included in the model as an explanatory variable together with the covariates $(\mathbf{x}, \mathbf{0})$ or $(\mathbf{x}, \mathbf{x})$, respectively. The duplication of data makes it possible to study possible interactions between covariates and failure types. If there are no ties present in data, then the Cox partial likelihood based on k failures is given by

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(c_i\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_1 + c_i\mathbf{x}_i'\boldsymbol{\beta}_2)}{\sum_{j\in R_i} \exp(c_j\beta_0 + \mathbf{x}_j'\boldsymbol{\beta}_1 + c_j\mathbf{x}_j'\boldsymbol{\beta}_2)}. \tag{2.53}$$

Considering a person with covariate vector $\mathbf{x}$, it can be seen that the hazard function for failure type I is $h_{01}(t)\exp(\mathbf{x}'\boldsymbol{\beta}_1)$, whereas the hazard function for failure type II is $h_{01}(t)\exp(\beta_0 + \mathbf{x}'\boldsymbol{\beta}_1 + \mathbf{x}'\boldsymbol{\beta}_2) = h_{02}(t)\exp(\mathbf{x}'\boldsymbol{\beta}_1 + \mathbf{x}'\boldsymbol{\beta}_2)$. Obviously, if $\mathbf{x} = 0$ the ratio of the baseline hazards functions of the two failure types equals to $\exp(\beta_0)$. However, this is no longer true when covariate vector $\mathbf{x} \neq 0$ is considered.

2.4 Frailty models

Frailty models can be considered as an extension of the proportional hazards models as they allow for addition of random effects to this kind of models [90]. In

addition to standard observed explaining variables, frailty represents a unobservable multiplicative effect on the hazard function. The idea behind frailty models is that individuals have different frailties, and that those who are most frail will experience the failure earlier than the others.

Two categories of frailty models can be considered. The first class of frailty models, called the univariate frailty models, can be used to model heterogeneity among individuals, i.e in terms of omitted variables, a univariate frailty model can be used when we assume the lack of measurements that vary within the group. The second class of frailty models, called the shared frailty models, takes into account correlated survival times and is used with multivariate survival data where the unobserved heterogeneity is shared among groups of individuals [44].

2.4.1 Univariate frailty models

It is no wonder that individuals are dissimilar in health sciences. At least for survival data, the sources of variability can be split into two groups: measurable risk factors and unknown covariates. Under the univariate frailty model, the heterogeneity represented by the missing information can be accounted for using the unobservable random variable, denoted as Z. Letting z_i be the frailty for the ith individual, then the hazard function at time t for individual i with vector of covariates $\mathbf{x}_i$, $i = 1, \ldots, n$, is given by $h(t, z_i, \mathbf{x}_i) = z_i h(t, \mathbf{x}_i)$, which is in the case of Cox model, equal to:

$$h(t, z_i, \mathbf{x}_i) = z_i h_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta}). \tag{2.54}$$

The identifiability property [32] implies that the frailty variable, Z, is assumed to have mean one and variance θ. In this way, $h(t, \mathbf{x}_i)$ can be seen as an average hazard rate of an individual i with vector of covariates $\mathbf{x}_i$ [1]. It follows immediately, that individuals with $z_i > 1$ will have an increased risk of failure whereas individuals with $z_i < 1$ are less frail and will survive longer given a certain covariate pattern. The univariate frailty model can be also formulated in terms of the conditional survival function which is, however, not observable:

$$S(t, z, \mathbf{x}) = \exp\left(-z \int_0^t h(u, \mathbf{x}) du\right) = \exp(-zH(t, \mathbf{x})). \tag{2.55}$$

When considering that the individual survival function conditional in the frailty can be written as $S(t, z, \mathbf{x}) = [S(t, \mathbf{x})]^z$, then the population survival function can be estimated by integrating over z. That is, if the frailty term has a probability density function $f(z)$, we can write the population survival function as:

$$S_\theta(t, \mathbf{x}) = \int_0^\infty [S(t, \mathbf{x})]^z f(z) dz. \tag{2.56}$$

The subscript θ is used to denote the dependence of the population survival function on the frailty variance θ. When considering the probability distribution of the unknown risks between individuals, Z, it comes reasonable to think that this is similar to the distribution of known risks. There are several probability distributions applied in frailty models but the following two are used most often: the gamma distribution and the log-normal distribution [90]. Considering the gamma distribution with mean one and variance θ, the density function of Z can be written as

$$f(z,\theta) = \frac{z^{1/\theta-1}\exp(-z/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}. \tag{2.57}$$

To estimate the nuisance parameter θ, we need to derive the likelihood function. If the frailties were observed, then the log likelihood based on the observed data would be:

$$\begin{aligned}\log L(\beta) = &\sum_{i=1}^{n} d_i \log(h_0(t_i)\exp(\mathbf{x}_i'\boldsymbol{\beta})) - H_0(t_i) z_i \exp(\mathbf{x}_i'\boldsymbol{\beta}) + d_i \log(z_i) \\ &+ \frac{1}{\theta}\log\left(\frac{1}{\theta}\right) + \left(\frac{1}{\theta}-1\right)\log(z_i) - \frac{z_i}{\theta} - \log\left(\Gamma\left(\frac{1}{\theta}\right)\right)\end{aligned} \tag{2.58}$$

This likelihood can be maximized using an EM algorithm according to [4], which was originally proposed in [70]. During first step, the value of θ is considered to be fixed, which enables us to ignore the last five terms of (2.58) in the first place. Initial estimates of regression coefficients vector, $\boldsymbol{\beta}$, and the cumulative baseline hazard function, $H_0(t)$, are obtained from the standard Cox model and Breslow's estimate, respectively.

Having the initial estimates of $\boldsymbol{\beta}$ and $H_0(t)$, then the following steps are iterated until convergence:

1. **The E step**: For univariate gamma frailty model, the expectation to be calculated is

$$\mathrm{E}(z_i|t_i, d_i, \mathbf{x}_i) = \frac{1+\theta d_i}{1+\theta\hat{H}_0(t_i)\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}, \tag{2.59}$$

where $\hat{H}_0(t_i)$ and $\hat{\boldsymbol{\beta}}$ are the actual estimates of $H_0(t)$ and $\boldsymbol{\beta}$, respectively.

2. **The M step**: During the M step we maximize

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \log(h_0(t_i)\exp(\mathbf{x}_i'\boldsymbol{\beta})) - H_0(t_i) z_i \exp(\mathbf{x}_i'\boldsymbol{\beta}) \tag{2.60}$$

using the standard Cox's partial likelihood with the expectation term $\mathrm{E}(z_i|t_i, d_i, \mathbf{x}_i)$ as an offset. The final estimates denoted as $\hat{H}_{0\theta}$ and $\hat{\boldsymbol{\beta}}_\theta$ are substituted into the

marginal log likelihood, $\log L_M$, after integrating out the frailties, i.e. into the following formula

$$\begin{aligned}\log L_M(\theta|\hat{H}_{0\theta}, \hat{\boldsymbol{\beta}}_\theta) = &\sum_{i=1}^{n} d_i \log(\hat{h}_{0\theta}(t_i)\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_\theta)) \\ &- \left(\frac{1}{\theta} + d_i\right)\log(1 + \hat{H}_{0\theta}(t_i)\theta\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_\theta)),\end{aligned} \tag{2.61}$$

where $\hat{h}_{0\theta}$ is estimated using the increments of $\hat{H}_{0\theta}$. The final estimate of θ is then derived numerically by maximizing (2.61).

2.4.2 Shared frailty models

There is often a situation, especially in medical data analysis, that data are clustered or correlated in some way, i.e. we assume unobserved heterogeneity shared by clusters of subjects. For example, this situation may arise in multicentric clinical studies or when we study failure times for samples consisting of natural families of individuals. In this model, all subjects within each group share a common frailty, each subject belongs to exactly one group, and frailties of different groups are independent. Difference between shared and univariate frailty models is in the assumption of how the frailty is distributed among the individuals.

Let the data of the ith individual, who is a member of the jth of J groups, follow a proportional hazards shared frailty model, then the hazard can be written as

$$h(t, z_{j(i)}, \mathbf{x}_i) = z_{j(i)} h_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta}). \tag{2.62}$$

where $j(i)$ denotes that the subject i is a member of the jth group, and $z_{j(i)} = z_j$ is the frailty for jth group. The individual z's are assumed to be independent and identically distributed according to some positive scale distribution with density function $f(z, \theta)$, having due to identifiability properties mean 1 and variance θ as already mentioned.

The estimation problem under this model can also be addressed using the EM algorithm [92, 59], with general framework proposed by Parner [72]. However, in 2000 Therneau, Grambsch & Pankratz have shown in [91] that the estimation under the gamma shared frailty model can be done using a penalized partial log likelihood, whose solution coincides with the solution given by the EM algorithm for any fixed value of θ. Moreover, they show that the Gaussian frailty models are also closely related to penalized models. Their findings were subsequently published also in [92].

The formulation of the shared frailty model as the penalized regression model is most easily derived using an alternative version of the frailty term

$$z_{j(i)} = \exp(\omega_{j(i)}). \tag{2.63}$$

Let us define a matrix of q indicator variables, $\mathbf{V}$, such that $v_{ij} = 1$ when subject i is a member of family j and 0 otherwise. Then the hazard function of the ith individual can be specified in another way:

$$h(t, z_{j(i)}, \mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\omega}), \tag{2.64}$$

where $\mathbf{v}_i$ is the column of $\mathbf{V}$ corresponding to the ith individual. Each subject is assumed to be a member of only one family. Then the penalized partial log likelihood can be written as

$$PPL = \log L(\boldsymbol{\beta}, \boldsymbol{\omega}) - g(\boldsymbol{\omega}; \theta) \tag{2.65}$$

where g is a penalty function which gives large values to "bad" values of $\boldsymbol{\omega}$, the parameter θ is a tuning constant, and $\log L(\boldsymbol{\beta}, \boldsymbol{\omega})$ is the standard Cox partial log likelihood of the form

$$\sum_{i=1}^{n} d_i \left\{ (\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\omega}) - \log \left[\sum_{k \in R_i} \exp(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{v}_k'\boldsymbol{\omega}) \right] \right\}. \tag{2.66}$$

The score equations, $\partial PPL/\partial\boldsymbol{\beta}$ and $\partial PPL/\partial\boldsymbol{\omega}$, need to be solved for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. As $\boldsymbol{\beta}$ is not included in the penalty function, then $\partial PPL/\partial\boldsymbol{\beta} = \partial \log L(\boldsymbol{\beta}, \boldsymbol{\omega})/\partial\boldsymbol{\beta}$, which implies that the score equations for $\boldsymbol{\beta}$ are equal to the score equations of a Cox model with $\mathbf{v}'\boldsymbol{\omega}$ as an offset term. Furthermore, we can define

$$\bar{v}_j(\boldsymbol{\beta}, \boldsymbol{\omega}, t) = \frac{\sum_{i \in R_t} v_{ij} \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\omega})}{\sum_{i \in R_t} \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\omega})}. \tag{2.67}$$

Then the differential of PPL according to ω_j can be written as

$$\frac{\partial PPL}{\partial \omega_j} = \sum_{i=1}^{n} d_i (v_{ij} - \bar{v}_j(\boldsymbol{\beta}, \boldsymbol{\omega}, t)) - \frac{\partial g(\boldsymbol{\omega}; \theta)}{\partial \omega_j}. \tag{2.68}$$

And the score equation for ω_j ($j = 1, \ldots, J$) can be written with the use of Breslow's estimator of the baseline hazard as

$$\frac{\partial PPL}{\partial \omega_j} = \sum_{i=1}^{n} (v_{ij} d_i - v_{ij} \hat{H}_0(t) \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\omega})) - \frac{\partial g(\boldsymbol{\omega}; \theta)}{\partial \omega_j} = 0. \tag{2.69}$$

The advantage of the penalized likelihood approach is that it can be fit using the Newton-Raphson algorithm.

2.5 Mixture cure model

Mixture survival models are useful when considering more than one parametric probability density to correctly describe the heterogeneity of data. Mixture survival models can be adopted to, for instance:

- Study diseases with a multiple stages, where time to failure in each stage is modelled with a different parametric distribution.
- Analyse several competing risks.
- Estimate a proportion of patients recovering after the treatment of the disease.

Last point is especially appealing in modelling population-based cancer data where patients can be (at least artificially) split into two groups: cured and uncured. Although clinically defined cure from cancer is never completely certain in an individual patient, mixture cure models are very useful for modelling the proportion of long term survivors among cancer patients on the population basis. Moreover, mixture models allow for incorporating and correcting for the background mortality, represented with the expected survival function, which is also of a great importance in analysing population-based cancer data [62].

Let us denote $h^*(t)$ and $S^*(t)$ the hazard function and the survival function of the general population, respectively; similarly, denote $h^R(t)$ and $S^R(t)$ the relative equivalents associated with the disease of interest; and finally, denote $h^U(t)$ and $S^U(t)$ the hazard function and the survival function of the uncured individuals, respectively. Moreover, let us consider π be the proportion of cured cases with respect to the specific cause of death and $1-\pi$ be the proportion of fatal cases that are bound to die of the specific cause of death. Then the mixture cure fraction model can be formulated as follows

$$S(t) = S^*(t)S^R(t) = S^*(t)(\pi + (1-\pi)S^U(t)), \tag{2.70}$$

with the excess mortality rate being in a form

$$h^R(t) = \frac{(1-\pi)f^U(t)}{\pi + (1-\pi)S^U(t)}. \tag{2.71}$$

2.5.1 Modelling the cure fraction

The aim of the mixture cure fraction model is often not only the estimation of π but also its modelling through covariates $\mathbf{x}$. There are two link functions mostly used in the cure fraction model [84]:

(i) *The identity link function*: $\pi(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. This link function causes the covariate effects to have relatively easy interpretation as they are represented in units of the cure fraction. However, use of the identity link function may cause problems in the boundary regions, i.e. for low or high cure fractions.

(ii) *The logistic link function*: $\log(\pi(\mathbf{x})/(1-\pi(\mathbf{x}))) = \mathbf{x}'\boldsymbol{\beta}$. Covariate effects are expressed as (log) odds ratios, and thus covariate effects have a similar interpretation to those in logistic regression.

Various probability distributions can be applied in the mixture cure models. For practical purposes, the Weibull, lognormal, and gamma distributions are mostly used. When considering a vector of covariates $\mathbf{x}$, the survival function of the uncured individuals, $S^U(t, \mathbf{x})$, can be written as

- $S^U(t, \mathbf{x}) = \exp(\lambda(\mathbf{x})t^{\gamma(\mathbf{x})})$ for the Weibull distribution, where λ and γ are a scale and a shape parameters, respectively.
- $S^U(t, \mathbf{x}) = 1 - \Phi(\frac{\log t - \mu}{\sigma})$ for the lognormal distribution, where Φ is the standard Normal distribution function.
- $S^U(t, \mathbf{x}) = 1 - \frac{\Gamma_t(a)}{\Gamma(a)}$ for the gamma distribution, where $\Gamma(a)$ is the gamma function and $\Gamma_t(a)$ is the incomplete gamma function, defined as $\Gamma_t(a) = \int_0^t x^{a-1}e^{-x}dx$.

All parameters of the mixture model can be estimated using the maximum likelihood approach. With (t_i, d_i) representing survival data of ith individual, the log likelihood function can be formulated as

$$\begin{aligned}\log L(\boldsymbol{\beta}) = &\sum_{i=1}^{n} d_i \log\left[h^*(t_i) + \frac{(1-\pi)f^U(t_i, \mathbf{x}_i)}{\pi + (1-\pi)S^U(t_i, \mathbf{x}_i)}\right] \\ &+ \log S^*(t_i) + \log\{\pi + (1-\pi)S^U(t_i, \mathbf{x}_i)\}.\end{aligned} \tag{2.72}$$

The expected survival, $S^*(t)$, does not depend on the model parameters and can be, thus, removed from the likelihood. However, the likelihood depends on the mortality rates of the general population through the expected hazard function, $h^*(t)$, which need to be estimated from the population mortality tables at either failure or censoring time of each individual [22]. Thus, the likelihood can be defined for any standard parametric distribution based on the probability density function $f^U(t)$ and survival function $S^U(t)$ characterising the uncured group. As in the Cox model, the inverse of the information matrix represents an asymptotic estimate of the covariance matrix of the parameter estimates.

Estimating number of patients potentially treated with anti-tumour therapy using population-based cancer registry data

3

The objective of this chapter is to present a new model for estimation of prevalence of patients requiring active anti-tumour therapy that should be accessible from population-based cancer registry data. The new model is an extension of a model already published in [27, 28] and [75]. The extension involves the way how the number of patients with terminal and non-terminal cancer is estimated. 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 budget impact of the treatment even more influencing than age at diagnosis. To document its applicability, the model has been applied on colorectal cancer data from the CNCR to model the number of potentially treated patients with colorectal carcinoma in the Czech Republic in 2011.

3.1 Introduction

Modern therapy introduces significant improvement in survival of cancer patients. However, the increasing cancer incidence and prevalence rates together with the cost of targeted anticancer therapy 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 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). However, irrespective of treatment applied, the cancer prevalence estimation is not an easy task. Since our primary interest is to estimate

the time interval prevalence, defined as the proportion (or number) of cancer patients ever diagnosed with cancer alive in the year of interest, it cannot be estimated directly from the population-based data due to time limited registration of the cancer cases, and has to be modelled. Several methods have been proposed for future cancer burden estimation based on different modelling strategies of which the back-calculation method combining parametric estimates of incidence and survival is most used [35, 76, 94, 88]. Other approaches include the calculation of individual log likelihoods of living with cancer [51], the application of Markov model [61] or the application of Bayesian model [5]. The generalization of completeness index method first introduced by Capocaccia & De Angelis [15] has also been applied [82].

3.2 Methodical concept of the model

The model 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 malignant 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.

3.2.1 Step I

Cancer prevalence is most frequently estimated using the cohort-specific exact age prevalence [94, 82] which denotes the probability that an individual with past or present history of cancer is alive at calendar time y and in the age range $[a, a+1)$. However, the exact age structure of prevalent individuals at time y is not always necessary from the financial budget impact point of view as the need for anti-tumour treatment should be judged rather by the presence or absence of cancer than by patient's age. In our model, we adopted an approach based on age-drift model emphasizing the period effects on cancer prevalence. It can be justified by the fact that the short-term predictions, which we focus here, are more likely to be influenced by period effects, such as cancer screening programmes or new treatment modalities, than by cohort effects. This is especially true in the

recent years when effects of cancer screening programmes and new treatment modalities on both cancer incidence and survival were reported for several cancers [7].

Considering the extent of cancer, s, as the main stratification factor for the estimation of cancer prevalence, let us 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:

$$P_s(y) = \sum_{a=1}^{m} P_{s,a}(y), \tag{3.1}$$

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 elements of (3.1) can be further formulated as the convolution of incidence and survival functions:

$$P_{s,a}(y) = \sum_{i=0}^{n} I_s(y-i,a)S_s(i,a), \tag{3.2}$$

where $I_s(y-i,a)$ 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 computation.

3.2.2 Step II

The purpose of the second step is to quantify the number of patients requiring active anti-tumour therapy on the basis of patients prevalent in the year of interest. Equation (3.2) can be easily split into two terms (assuming newly diagnosed patients being prevalent in the year of interest and thus making $S_s(0,a) = 1$) as follows:

$$P_{s,a}(y) = \sum_{i=0}^{n} I_s(y-i,a)S_s(i,a) = I_s(y,a) + \sum_{i=1}^{n} I_s(y-i,a)S_s(i,a). \tag{3.3}$$

First term on the right-hand side of (3.3) represents the newly diagnosed patients whereas the second one stands for the patients diagnosed in the past and alive in the given year. Correcting the first term of (3.3) 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 (3.3) 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:

$$P^*_{s,a}(y) = I_s(y,a)\delta_s(y,a) + \sum_{i=1}^{n} I_s(y-i,a)S_s(i,a)R_s(i,a)\delta_s(y,a), \tag{3.4}$$

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.

In a simplified way, the $R_s(i,a)$ can be further specified in the stage-specific manner 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; denoted as $R^1_s(i,a)$) or terminal (leading to death in the year of interest; denoted as $R^2_s(i,a)$), which further determine the way of patient's treatment. In the former case the patient is assumed to be treated in a similar way as in the time of primary diagnosis which implies the patient stays in the pool of patients prevalent in particular stage s. In the second case the patient is assumed to be treated with generalized disease which implies the patient moves from the prevalence of stage s (I+II or III) to prevalence of stage IV. Schematically, the $R_s(i,a)$ can be expressed as:

$$R_s(i,a) = R^1_s(i,a) + R^2_s(i,a);\ s = \text{I+II, III, IV}, \tag{3.5}$$

where $R^1_s(i,a)$ and $R^2_s(i,a)$ reflects the age and stage-specific probabilities of non-terminal and terminal cancer recurrence, respectively, conditional on being alive i years from diagnosis.

The simplifying assumption of only two forms of cancer recurrence has again the motivation in financial aspects of cancer care. The separation of patients with terminal cancer recurrence is needed for the treatment of generalized disease is much more costly than the treatment of non-terminal disease, especially when considering the contemporary improvements in targeted anti-tumour therapy. However, the remaining clinical stages (I+II and III) can be with respect to financial burden perceived as much more similar.

Expanding (3.4) in a way given above, i.e. splitting the $R_s(i,a)$ term according to (3.5) and moving the patients suffering from terminal cancer recurrence to prevalence of stage IV, lead to final formulation of the age- and stage-specific prevalence of patients requiring active anti-tumour therapy as follows:

$$
\begin{aligned}
P^*_{s,a}(y) &= I_s(y,a)\delta_s(y,a) + \sum_{i=1}^{n} I_s(y-i,a)S_s(i,a)R^1_s(i,a)\delta_s(y,a);\ s = \text{I+II, III,} \\
P^*_{\text{IV},a}(y) &= I_{\text{IV}}(y,a)\delta_{\text{IV}}(y,a) + \sum_{i=1}^{n} I_{\text{IV}}(y-i,a)S_{\text{IV}}(i,a)(R^1_{\text{IV}}(i,a) + R^2_{\text{IV}}(i,a))\delta_{\text{IV}}(y,a) \\
&\quad + \sum_{s=\text{I+II,III}} \sum_{i=1}^{n} I_s(y-i,a)S_s(i,a)R^2_s(i,a)\delta_s(y,a).
\end{aligned}
\tag{3.6}
$$

The components of (3.6) are further specified in detail in the following sections.

3.2.3 Model for cancer incidence

Extrapolation of past incidence trends is a standard way how to obtain future incidence rates in cancer prevalence modelling [30]. The cancer incidence model needed for specification of (3.6) is based on the age, period and cohort model [52] which is clearly the most often applied model for cancer incidence [8]. In the proposed model, the age-drift Poisson regression models are used employing two different link functions. Different models for increasing and decreasing incidence trends are utilised to prevent explosive exponential growth for increasing trends and negative values for decreasing trends [29]. It has been shown previously that the age-drift Poisson regression model is easy to implement and gives reasonably accurate predictions [66]. Stage-specific models can be fit as follows:

$$\log(\mathrm{E}(I_s(y,a)/n_{y,a})) = \alpha_{s,a} + \beta_s y, \tag{3.7}$$

where y denotes the calendar year of interest, a is the age category and $n_{y,a}$ is the number of person-years at ath age category and yth calendar year. The use of $n_{y,a}$ introduces adjustment for the changing demographic structure of the considered population. Subsequently, overall slope of the stage-specific model (drift parameter β_s) is assessed and, in case of increasing overall trend in number of cancer cases, the model using identity link is employed:

$$\mathrm{E}(I_s(y,a)/n_{y,a}) = \alpha_{s,a} + \beta_s y. \tag{3.8}$$

In case of different slopes in individual age categories (as assessed by the likelihood ratio test), models with age-specific drift parameter are utilised. Model with log link function is used for fitting of the decreasing overall trend:

$$\log(\mathrm{E}(I_s(y,a)/n_{y,a})) = \alpha_{s,a} + \beta_{s,a} y, \tag{3.9}$$

whereas model with identity link function is used for fitting of the increasing overall trend:

$$\mathrm{E}(I_s(y,a)/n_{y,a}) = \alpha_{s,a} + \beta_{s,a} y. \tag{3.10}$$

It is also possible to supplement point projections with prediction intervals using standard methodology [47].

3.2.4 Survival estimates

Second component needed for computation of (3.6) are the age- and stage-specific survival rates. Unlike other models for cancer prevalence [94, 82], the proposed model is not based on birth cohort specific prevalence and thus the effect of general mortality is not implied in (3.2). Thus, we have to calculate and apply the estimates of standard cumulative observed survival rates. The life-table method is employed as a standard method to estimate observed survival rates using data from population-based registries [6, 18]; this method processes one-year periods, and its accuracy is therefore not very much affected by detailed quality of records within the registry. Considering the fact that significant changes regarding survival rates of cancer patients can be observed in time, the estimates of x-year survival rates are derived using the principle of the so-called moving window that will be further described in detail.

In this methodology, the x-year survival rates are estimated successively, using the cohort analysis of patients diagnosed in a five-year time intervals (e.g. cohorts of patients diagnosed in years 2003–2007, 2002–2006, etc.). Each of these cohorts provides information on one-year survival rate to x-year survival rate, where x is the number of years from the start of the time interval to the last available date reported in the population-based registry; 31. 12. 2007 may serve as an example. Then, for example, the cohort of patients diagnosed between 1995 and 1999 provides information needed to calculate the 1-year, 2-year, ... 13-year survival rates (x = 13, which is the number of one-year periods between 1.1.1995 and 31.12.2007, as explained above).

However, cohorts of patients diagnosed many years ago are not used to estimate short-term survival rates (such as 1-year, 2-year etc.), as the resulting estimates should be obviously biased downwards. For this reason, calculation of x-year survival rates is only performed on patient cohorts in which x-year survival rate can be reliably estimated, and which were diagnosed as recently as possible. In other words, patients diagnosed in 2003–2007 will contribute to the estimate of 1-year to 5-year survival rates, patients diagnosed in 2002–2006 will contribute to the estimate of 2-year to 6-year survival rates, etc. The width of interval defining one patient cohort was set to five years, as this is a standard width

used in population-based survival analyses [6]. The estimate of $S_s(i,a)$, needed for calculation of (3.6), can be then expressed as follows:

$$\hat{S}_s(i,a) = \text{median}(\hat{S}_s^{Y-i+1}(i,a); \hat{S}_s^{Y-i+2}(i,a); \hat{S}_s^{Y-i+3}(i,a); \hat{S}_s^{Y-i+4}(i,a); \hat{S}_s^{Y-i+5}(i,a)) \tag{3.11}$$

where Y is the maximum year of follow-up of the population-based registry, $Y-i+l(l=1,...,5)$ are the years defining the upper limit of the 5-year time period for the patient cohort selection and $\hat{S}_s^{Y-i+l}(i,a)$ is the corresponding age- and stage-specific estimate of i-year survival rate calculated on that cohort. For example, there is only one available cohort providing information on 1-year survival rate and that is the cohort of patients diagnosed in 2003–2007, whereas for 2-year survival rate two cohorts of patients, 2003–2007 and 2002–2006, provide relevant information. Another example could be the estimation of 7-year survival rate which would be based on cohorts of patient diagnosed in 2001–2005, 2000–2004, 1999–2003, 1998–2002 and 1997–2001.

3.2.5 Non-terminal cancer recurrence rates

Since the precise information on time of cancer recurrence is barely available in population-based cancer registries, the rationale behind the estimation of $R_s^1(i,a)$ and $R_s^2(i,a)$ functions is to use surrogate parameters with direct association to the probability of cancer recurrence. Considering non-terminal cancer recurrence as the first case, $R_s^1(i,a)$ is estimated using the information on the patient's health status and non-symptomatic anti-tumour therapy applied during the follow-up period which ensues the time of diagnosis, i.e. which ensues the year of diagnosis when the patient is a part of cancer incidence. This approach simply assumes that record on other than symptomatic therapy in a particular year after diagnosis indicates that the patient is treated due to objective reason, i.e. due to the return of cancer. However, as the $R_s^1(i,a)$ function refers to non-terminal cancer recurrence, there is an additional condition needed and that is the patients have to survive up to the end of the particular year of interest, i.e. the cancer recurrence should not be terminal in that year. As the date of cancer patient's health status and treatment assessment, respectively, in the follow-up period cannot be measured with required precision, we need to adopt for $R_s^1(i,a)$ function estimation a method working with grouped lifetime data, i.e. utilising the life-table principle. More specifically, we suggest a nonparametric smoothing technique to be used for an initial hazard estimate provided by the life-table method. Let p be the number of knots chosen for the smoothing and $h_s(j,a)$ represent the life-table hazard estimate for the jth time interval. Then, generally, the hazard rate estimate is of the form

$$\hat{R}_s^1(i,a) = \sum_{j=1}^{p} c_j(i)\hat{h}_s(j,a), \tag{3.12}$$

where $\sum_{j=1}^{p} c_j(i) = 1$ for each time i. The values $c_j(i)$ thus represent weights specified by the smoothing method, the resulting hazard estimate at time i is a weighted average of the hazard rates given by the life-table method. The choice of a particular smoothing technique is limited by the quality of cancer registry data and the number of follow-up time points at which the information on patient is recorded but, in general, for the estimation based on grouped data, the locally weighted least squares method is recommended [97].

Since all smoothing techniques have problems with proper estimation in the boundary regions, which in case of $R_s^1(i,a)$ function could have lead to underestimation of the hazard in the first time interval, we suggest the probability of non-terminal cancer recurrence in first year after diagnosis, i.e. $\hat{R}_s^1(1,a)$, to be fixed at the value of the life-table hazard estimate for the first time interval, i.e. $\hat{h}_s(1,a)$.

As the purpose of $R_s^1(i,a)$ is to cover the risk of first non-terminal cancer recurrence after primary diagnosis, the information on anti-tumour therapy used for its identification is derived only from the first time interval at which the non-symptomatic therapy is recorded even if the patient can have recorded the therapy at several consecutive time intervals.

3.2.6 Terminal cancer recurrence rates

As for the terminal cancer recurrence, the $R_s^2(i,a)$ function can be estimated using the information on cancer as the cause of death recorded in the population-based registry. The approach we have adopted is based on the assumption that nobody can die from cancer without passing through the phase of generalized disease, i.e. even the patient diagnosed primarily in stage I or II can be thought as treated with distant disease in the future when cancer is recorded as his cause of death. The $R_s^2(i)$ function thus represents the excess mortality of the cancer and can be thus specified using the relative survival function or, more specifically, using the underlying excess hazard rate. Both these quantities are derived using the mixture cure survival model adjusted for background mortality [22] as it allows for incorporating covariates. With s representing the stage of disease and a representing the ath age category, the mixture cure survival model is given by

$$S_s(i,a) = S^*(i,a)S_s^R(i,a) = S^*(i,a)\{\pi_s(a) + (1-\pi_s(a))S_s^U(i,a)\}, \tag{3.13}$$

where $S^*(i,a)$ is the expected survival for age class a easily accessible from the national mortality statistics, $S_s^R(i,a)$ is the stage-specific relative survival function of the whole cohort of patients, $\pi_s(a)$ is the stage-specific proportion of patients cured from cancer and $S_s^U(i,a)$ is the stage-specific relative survival function of the uncured patients. The latter can be specified parametrically using standard distribution functions, e.g. Weibull or lognormal, the parameters of which can be further modelled considering various covariates as the age at diagnosis or period of diagnosis. Letting $f_s^R(i,a)$ and $f_s^U(i,a)$, respectively, be the probability density function associated with $S_s^R(i,a)$ and $S_s^U(i,a)$, respectively, the excess hazard rate of the whole cohort of patients is given by

$$h_s^R(i,a) = \frac{f_s^R(i,a)}{S_s^R(i,a)} = \frac{(1-\pi_s(a))f_s^U(i,a)}{\pi_s(a) + (1-\pi_s(a))S_s^U(i,a)}. \tag{3.14}$$

It should be noted that the mixture cure survival model-based estimates of $h_s^R(i,a)$ cannot be immediately applied with respect to index i for the $R_s^2(i)$ function approximation as they refer to instantaneous risk of cancer recurrence whilst we need to estimate the risk integrated over one year time period. However, the estimate of $R_s^2(i)$ can be easily calculated using the elementary formula

$$\hat{R}_s^2(i,a) = \int_i^{i+1} \hat{h}_s^R(t,a)dt = -\log(\hat{S}_s^R(i+1,a)) + \log(\hat{S}_s^R(i,a)). \tag{3.15}$$

3.2.7 Modelling proportion of patients treated with anti-tumour therapy

Estimates of the proportions of patients treated with anti-tumour therapy are also needed for calculation of (3.6) as they represent correcting factors reflecting the patients' health status and influence the probability of anti-tumour therapy administration. The estimated proportions of treated patients in time can be derived from the population-based registry and further extrapolated back and forward in time using simple logistic regression model:

$$\text{logit}(\text{E}(\delta_s(y,a))) = \alpha_(s,a) + \beta_s y. \tag{3.16}$$

There is a strong association between age, stage and patients' health status suggesting that separate logistic models should be fit for individual clinical stages giving us age- and stage-specific estimates of proportion of patients to be treated in the years to come.

It should be noted that we assume in (3.6) that the probability of treatment for patients diagnosed in the past is the same during the follow-up as in the time of

diagnosis. That means we assume the probability of treatment for cancer recurrence to be age-specific with respect to age at diagnosis instead of age at cancer recurrence. However, since the risk of cancer recurrence is mostly apparent during first few years following the diagnosis, the bias introduced with this assumption should be rather small.

3.3 Modelling the colorectal cancer in the Czech Republic

3.3.1 Data source

Proper prevalence estimates can be only calculated if high-quality data from registries with long-standing registration and sufficient follow-up are used [94, 33]. The Czech Republic disposes of representative population-based data that constitutes a high-quality basis for such analyses [67]. The Czech National Cancer Registry (CNCR) covers whole population of the Czech Republic since 1977 and its 100 % coverage means that this database is fully covering of the Czech population. Until December 2007, there was over 1.5 million cancer cases recorded in the CNCR.

3.3.2 Results

To demonstrate the applicability of the presented model and its potential usefulness in financial budget impact analyses, data on 123,562 primary colorectal cancer cases (ICD-10 codes C18–C20) diagnosed and staged in 1982–2007 were used to estimate the number of patients requiring active anti-tumour therapy in the Czech Republic in 2011. Data on cases diagnosed in 1977–1981 were omitted due to the lack of classification system of clinical stages. All patients diagnosed by death certificate only (DCO) or at autopsy were excluded from the analysis. All estimates have been derived in age- and stage- specific manner considering four age categories: 15–49 years, 50–64 years, 65–79 years and 80+ years; and three categories representing clinical stages: stage I+II, stage III and stage IV. In survival modelling utilised for risk of terminal cancer recurrence estimation, the effect of time period of diagnosis was considered and following time periods were used: 1982–1989, 1990–1994 and 1995–2007. These periods were chosen to reflect the main developmental stages of the Czech health care system and TNM classification system. All computations were performed using STATA 10.0 software [85].

As a very first part of the estimation process, an incidence model was set up. We have decided to use a prediction base of 10-year length since the models based

on shorter prediction base seem to perform better because they don't include outdated information when estimating the drift component [66]. To obtain as long time series of incidence figures as possible, two prediction directions were included in the analysis. Recent data corresponding to observation period 1998–2007 were employed to predict incidence rates for the period 2008–2011, while older data, corresponding to observation period 1982–1991, were used to project incidence trends to the past, covering the time period 1970–1981. Colorectal cancer incidence rates per 100,000 people available for 2007 from CNCR and projections of colorectal cancer incidence rates per 100,000 people in the Czech Republic for 2008–2011 according to the clinical stage of primary tumour are given in Table 3.1. The point estimates are accompanied with 95% confidence and prediction intervals, respectively, using standard methodology [47]. In absolute numbers, this corresponds to 2,909 newly diagnosed patients with stage I or II colorectal cancer in 2011 in the Czech Republic, 1,742 newly diagnosed patients with stage III, and 1,595 newly diagnosed patients with generalised disease.

Table 3.1: Colorectal cancer incidence rates per 100,000 people for 2007 (last available year from the CNCR) and estimated colorectal cancer incidence rates per 100,000 people for 2008–2011 according to the clinical stage of primary tumour.

Cancer stage	2007	2008	2009	2010	2011
Stage I+II	25.8 (24.8–26.8)	27.3 (26.1–28.5)	27.3 (26.0–28.6)	27.3 (26.0–28.7)	27.5 (26.1–28.9)
Stage III	14.0 (13.3–14.7)	14.9 (13.7–16.0)	15.4 (14.2–16.6)	15.9 (14.6–17.2)	16.5 (15.1–17.8)
Stage IV	14.6 (13.9–15.4)	14.9 (14.1–15.7)	15.0 (14.1–15.8)	15.0 (14.1–15.8)	15.1 (14.2–16.0)
Total	54.4 (53.0–55.8)	57.1 (55.6–58.5)	57.6 (56.2–59.1)	58.2 (56.7–59.6)	59.1 (57.6–60.6)

To complete the 2011 cancer prevalence estimates, the cumulative survival rates were derived by the moving window cohort analysis for individual age- and stage- specific categories. However, the follow-up of the colorectal cancer patients is limited to 1982–2007 time series and thus the CNCR data provide only the information on 1-year, 2-year, ... 26-year survival probabilities. Since the cancer incidence has been projected back to 1970, there was a need for projecting the survival rates to get also the 27-year, 28-year, 29-year, ... 41-year survival rates. The missing survival rates have been extrapolated from the observed survival rates using the exponential survival model. The observed as well as the projected survival rates are given in Table 3.2. There can be easily seen a quick decrease in survival probabilities in older cancer patients with higher extent of the disease. The long term survival is above 10 % only for patients diagnosed in younger age and lower clinical stage.

However, it has to be stressed out that the usability of the survival rates calculated by the moving window cohort analysis is limited only to the proposed model as the rates are not standardized for background mortality and as such are

not usable for comparison with other populations or evaluation of time trends.

Table 3.2: Age- and stage- specific survival rates derived by the moving window cohort analysis.

	Clinical stage I + II				Clinical stage III				Clinical stage IV			
	15-49 years	50-64 years	65-79 years	80+ years	15-49 years	50-64 years	65-79 years	80+ years	15-49 years	50-64 years	65-79 years	80+ years
1-year	97%	93%	85%	68%	91%	90%	79%	59%	60%	51%	37%	18%
5-year	81%	76%	60%	34%	57%	52%	38%	22%	15%	11%	8%	4%
10-year	70%	55%	34%	11%	37%	32%	19%	8%	8%	5%	3%	1%
15-year	51%	37%	15%	4%	29%	20%	11%	2%	5%	4%	2%	1%
20-year	44%	25%	6%	1%	24%	14%	4%	0%	3%	2%	1%	0%
30-year	30%	12%	1%	0%	13%	5%	1%	0%	1%	0%	0%	0%
40-year	20%	5%	0%	0%	7%	2%	0%	0%	1%	0%	0%	0%

Figure 3.1 shows the observed values and the predicted value of colorectal cancer interval prevalence per 100,000 people in the Czech Republic according to clinical stage of primary tumour. The point estimate is accompanied with the 95% confidence intervals calculated according to [17]. There can be seen a good agreement of the predicted value of the 2011 interval prevalence and the observed values up to year 2007 in all considered stage groups. In 2011, colorectal cancer prevalence is estimated to be 280.4 cases per 100,000 people for patients primarily diagnosed in stage I or II (ranging between 277.4 and 283.6 per 100,000), 95.3 cases per 100,000 people for patients diagnosed in stage III (ranging between 93.5 and 97.2 per 100,000) and 37.2 cases per 100,000 people for patients diagnosed in stage IV (ranging between 36.0 and 38.3 per 100,000). In total, 412.9 staged colorectal cancer patients per 100,000 people are to be prevalent in 2011 which introduces more than 17 % increase with respect to year 2007. Represented as an absolute number, more than 43,000 patients with completed tumour diagnostics are likely to be prevalent with colorectal cancer in 2011 in the Czech Republic.

To proceed from cancer prevalence estimates to number of patients that are likely to require active anti-tumour therapy in the future, the rates of cancer recurrence were estimated.

The non-terminal cancer recurrence rates were derived with respect to clinical stage, age and period of diagnosis from the CNCR Follow-up Reports on Malignant Neoplasms by the life-table method based on one-year time intervals and smoothed afterwards using kernel-weighted polynomial regression [97]. As a smoothing function we used quadratic polynomial with Gaussian kernel for weighting. These estimates in a standard way refer to the conditional probability of suffering from cancer recurrence in a particular year y after diagnosis given that a patient has survived up to the beginning of the year y and, in addition, also that a patient has survived up to the end of the year y, i.e. the cancer recurrence was not terminal in the year y.

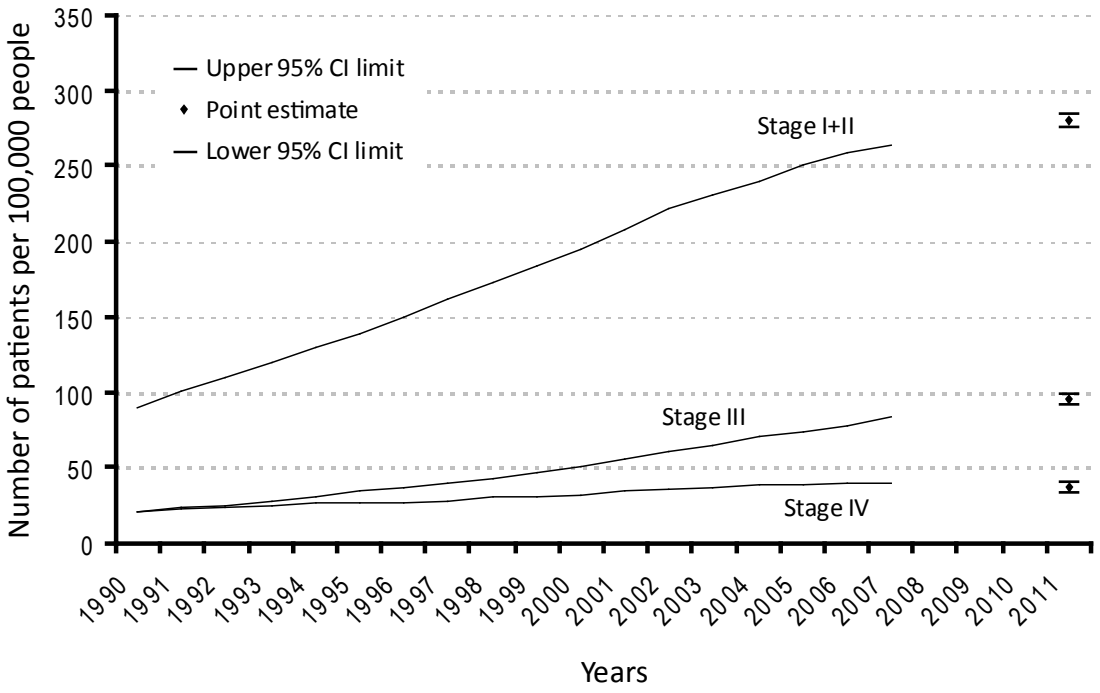

Figure 3.1: Observed and predicted values of colorectal cancer prevalence in the Czech Republic per 100,000 people according to clinical stage of primary tumour.

The terminal cancer recurrence rates were derived using the stage-specific mixture cure relative survival model considering age category and period of diagnosis as covariates. The model fit was assessed visually according to comparison with standard Hakulinen relative survival estimates [45]. Finally, the Weibull distribution was used to model the survival distribution of the uncured patients in stage IV cancer whereas the mixture of two Weibull distributions were used in stage I+II and stage III cancer.

Figure 3.2 shows the estimated rates of non-terminal and terminal cancer recurrence, respectively, with respect to clinical stage and age category in ten years after the first completed year from diagnosis; the estimates correspond to the most recent time period, 1995-2007. The risk of non-terminal cancer recurrence is most noticeable in four years after the first completed year from diagnosis and then gradually decreases towards zero in all stages. However, still there can be seen remarkable differences among the clinical stages with stage I+II being the category with the lowest risk of non-terminal cancer recurrence and stage IV being the category with the highest risk of non-terminal cancer recurrence. The pattern of terminal cancer recurrence rates vary greatly with clinical stage; in stage I+II, we can see the recurrence rates being below 5 % all the time giving the evi-

dence of good perspectives of patients diagnosed with less advanced disease. In stage III, the terminal recurrence rates show a stable level in three years after the first completed year from diagnosis with a slow decrease afterwards. The only difference is the oldest age group showing higher risk of dying from colorectal cancer after first year from diagnosis and steeper decrease of the recurrence rate afterwards. In contrast, the terminal recurrence rate of stage IV show very high risk of dying from colorectal cancer even after first completed year from diagnosis reaching initially 50 % in all age groups. The subsequent decrease in this rate is rather slow and age-dependent resulting in an appreciable risk even after several years from diagnosis. It should be noted that the cancer recurrence rates of the oldest age group are almost in all cases lower than the rates of remaining age groups. This can be explained with much higher mortality of the oldest patients during the first year after diagnosis (see Table 3.2) which causes only the fittest of this age group to survive and thus remain at-risk also after the first year from diagnosis.

Last component needed for the estimation of number of patients requiring active anti-tumour therapy are the age- and stage-specific proportions of patients treated with an anti-tumour therapy in 2011. Time trend analysis of the CNCR data has shown similar increasing pattern of these proportions in all age groups, whereas with respect to stage the analysis has shown stable proportion in stages I and II, slight time-dependent increase in stage III and sharp increase in anti-tumour therapy of generalized disease. Following portions of treated patients were predicted for 2011: 97.3 %, 98.0 %, 96.4 %, and 88.0 % of patients, respectively, diagnosed with stage I or II in the considered age groups: 15-49 years, 50-64 years, 65-79 years, and 80+ years, respectively. Similar figures were observed in patients diagnosed with stage III: 99.5 %, 99.2 %, 98.3 %, and 94.7 %, respectively. Fewer patients with stage IV colorectal cancer may benefit from anti-tumour therapy due to poor health condition, particularly in the oldest age group, meaning that only 92.3 %, 88.9 %, 80.3 % and 57.9 % of patients, respectively, will be probably treated for metastatic colorectal cancer in 2011.

The estimated numbers of patients requiring active anti-tumour therapy for colorectal cancer in the Czech Republic in 2011 are presented in Table 3.3 and are accompanied with 95% confidence intervals calculated according to [17]. The results are given in two ways: Firstly, the numbers of potentially treated patients with respect to clinical stage at the time of diagnosis are shown in the upper part of Table 3.3. These numbers reflect the anticipated cancer burden generated by original clinical stages in three considered levels, i.e. newly diagnosed patients never before treated for cancer, patients assumed to be treated for non-terminal cancer recurrence and patients assumed to be treated with terminal disease recurrence. Secondly, in the lower part of Table 3.3 the numbers of patients potentially

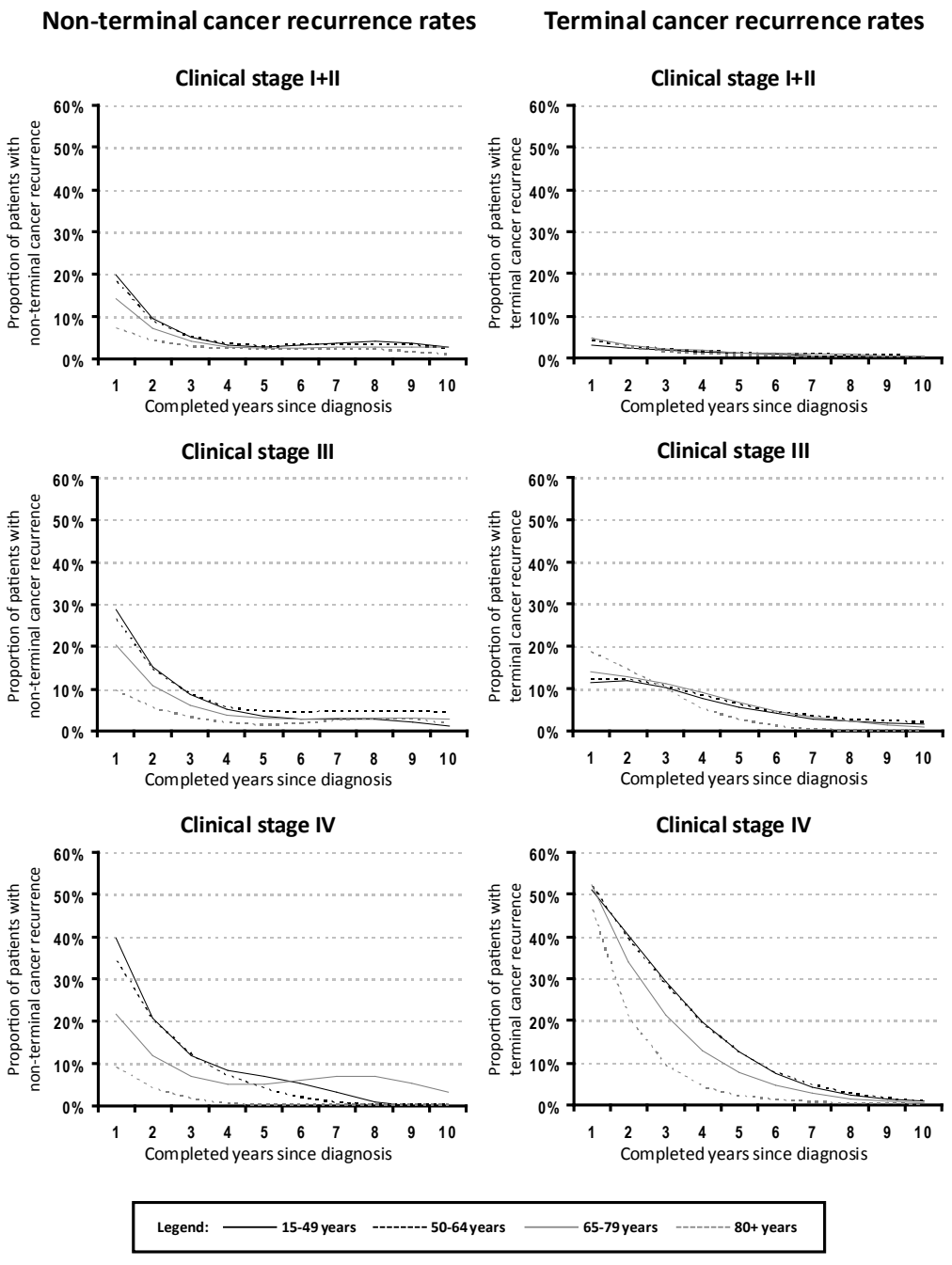

Figure 3.2: Stage- and age-specific estimates of non-terminal and terminal cancer recurrence rates in first ten years after diagnosis; the estimates correspond to the more recent time period, 1995-2007.

Table 3.3: Estimated numbers of patients requiring active anti-tumour therapy for colorectal cancer in the Czech Republic in 2011 according to clinical stage at diagnosis and putative stage in 2011 indicating actual extent of the disease.

Stage	Patients potentially treated in 2011 according to stage at diagnosis			
at diagnosis	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total number of patients
Stage I+II	2778 (2676-2883)	1058 (995-1124)	386 (348-426)	4222 (4096-4351)
Stage III	1709 (1629-1792)	620 (572-671)	464 (423-508)	2793 (2690-2899)
Stage IV	1281 (1212-1353)	232 (203-263)	469 (428-513)	1982 (1896-2071)
All stages	5768 (5620-5919)	1910 (1825-1998)	1319 (1249-1392)	8997 (8812-9185)
Putative	Patients potentially treated in 2011 according to putative extent of disease			
stage in 2011	Newly diagnosed patients	Non-terminal cancer recurrence	Terminal cancer recurrence	Total number of patients
Stage I+II	2778 (2676-2883)	1058 (995-1124)	-	3836 (3716-3959)
Stage III	1709 (1629-1792)	620 (572-671)	-	2329 (2235-2426)
Stage IV	1281 (1212-1353)	232 (203-263)	1319 (1249-1392)	2832 (2729-2938)
All stages	5768 (5620-5919)	1910 (1825-1998)	1319 (1249-1392)	8997 (8812-9185)

treated with the anti-tumour therapy are shown according to the expected disease extent of these patients in 2011. These values reflect the actual staging of potentially treated patients that should be accounted for in financial planning for oncology care. It can be seen that the main difference between the upper and lower parts of Table 3.3 is in the number of patients treated for terminal cancer recurrence. As mentioned before, we assume that nobody can die from cancer without passing through the phase of generalized disease, i.e. if the patient is to be treated for terminal cancer recurrence, he is assumed to be treated in stage IV.

In total, almost 9,000 colorectal cancer patients in different stages of the disease are predicted for anti-tumour therapy administration in the Czech Republic in 2011. This number constitutes indispensable financial burden that need to be accounted for in financial planning in health care. Moreover, almost one third of these patients are predicted as treated in stage IV which represents the worst condition from both health and financial perspectives.

3.4 Discussion

When considering the financial aspects of present-day anti-tumour therapy, specific estimate of the prevalence of patients requiring active anti-tumour treatment can be of interest besides the standard estimates of cancer incidence and prevalence. In this chapter, a statistical method was proposed that may provide such estimates utilizing the population-based cancer registry data. The estimation pro-

cess respecting the extent of patient's disease is divided in two subsequent steps. In the first step, the amount of patients ever diagnosed with cancer and alive in given time interval in the population of interest is quantified, i.e. time interval prevalence of cancer patients is estimated. Then, in the second step, the prevalence estimate is corrected for patients not treated with anti-tumour therapy from whatever reason (cure for cancer, treatment contraindication, very high age, patient's refusal to treatment, advanced disease).

All estimates are derived solely from population based data using the back-calculation procedure that combines the observed and modelled cancer incidence, survival and hazard estimates. For practical purposes, the model was proposed in discrete time fashion assuming one calendar year as time unit. The patient is thus assumed to be prevalent in the year of interest even if that patient dies during the one-year period for he should be regarded as potential for the anti-tumour therapy administration. This assumption is motivated by the financial planning process in health care which is mainly based on one-year time periods. Moreover, the proposed model was formulated with respect to the extent of disease as well as the age at diagnosis. The effort to estimate the stage-specific cancer prevalence is a challenging task with little evidence on this subject being found in the literature [42]. However, in contrast to other cancer prevalence models [94, 82], the age structure of patients prevalent in the year of interest cannot be estimated with this model for it is not formulated on the basis of birth cohorts. An argument can be that the age structure of individuals prevalent in specific year is not of main interest of the health care payers and providers as the need for anti-tumour treatment should be judged mainly by the presence or absence of cancer and not merely by patient's age.

Observed stage-specific colorectal cancer incidence was utilised for extrapolation using classic Poisson regression model. Simple age-drift model was used to prevent overfitting of observed time series, which might be introduced by including complex terms in the model. To obtain maximum coverage of prevalent cancer cases, backward prediction of cancer rates in pre-registration period was used beyond the standard prediction of future rates. The observed survival rates needed for the estimation process are derived using a moving window principle defining cohorts of patients having similar year of diagnosis that are further used for the calculation of survival rates. Unlike the well-known period analysis which is standard in population survival analyses, this method does not provide up-to-date estimates of survival rates relevant only for newly diagnosed patients but provide survival rates relevant for modelling short-term cancer prevalence.

The issue of stage-specific cancer prevalence estimation can be considered controversial due to nontrivial association between the stage at diagnosis and the gradual progress of cancer during the follow-up period. Cancer recurrence

rates of patients diagnosed in the past and living in the year of interest are thus by all means the most appealing and most arguable components of the model, especially when considering its estimation from population-based cancer registry data. The cancer patient's status is often monitored in the pre-defined time intervals, but the precise information on time of cancer recurrence can be rarely available. However, this component cannot be omitted and has to be estimated. Moreover, the estimate coming from other than population-based databases can also lead to biased results due to non-representativeness of the underlying set of patients.

There are two types of cancer recurrence considered in the model, terminal rates and non-terminal rates, each of them being estimated differently. The non-terminal recurrence rates are estimated from the records on other than symptomatic therapy during follow-up after the diagnosis indicating objective reason for anti-tumour treatment administration. The terminal recurrence rates are derived using mixture cure fraction survival model [22] accounting for background mortality representing causes of death other than cancer. This model assumes the set of patients to be a mixture of fatal and cured patients and even if this assumption of being cured of cancer at the date of diagnosis is not plausible from the biological perspective, this model has been shown to be relevant for modelling cancer survival and hazard functions [62, 84]. Possible confusion, however, may be introduced to the estimation of non-terminal recurrence rates as the proposed model consider only the first time the anti-tumour therapy administration is recorded after completion of one year from diagnosis. The recurring cancer relapses and progressions cannot be exactly monitored using population-based data for the insufficient detail of clinical information available from these data. Elaborate analysis of clinical data of hospital-based nature should be processed to get complete insight into cancer recurrence in time.

To document the applicability of the proposed model, it has been applied on colorectal cancer data coming from the CNCR to predict the number of patients requiring active anti-tumour therapy for colorectal carcinoma in the Czech Republic in 2011. A good fit was obtained from models for future cancer incidence prediction of the considered stage groups: I+II, III, IV. Age specific incidence slopes were included in model of disease of clinical stages I+II (significant decrease at age under 80 years, non-significant increase over 80 years) and III (significant increase at age over 50 years, non-significant increase in youngest age group). Significant decreasing trend common for all age groups was observed in clinical stage IV. Significant extra-Poisson variation was observed only in stage III model (estimated dispersion factor 1.60).

The age-standardized incidence rate of colorectal cancer in the Czech Republic is among the highest in Europe [34]. Due to unfavourable development in demo-

graphic structure, the number of colorectal cancer cases is increasing, no matter the mentioned decrease in most of age-stage-specific rates. Recent increase in number of colorectal cancer cases has been observed in other Central and Eastern Europe cancer registries [71]. Increase in colorectal cancer incidence was also predicted in Italian study [41]. The smooth trends in colorectal cancer incidence rates could be disturbed by cancer control programmes, including primary prevention campaigns or colorectal cancer screening programme. Colorectal cancer screening programme is able to affect incidence of the disease [64]. Extent of overestimating the future rates might therefore be used for evaluation of success of such programmes [46]. However, low compliance to available screening regimen has been precluding impact on colorectal cancer incidence in the Czech Republic so far. No special factors regarding cancer prevention and screening were therefore included in the model.

The observed survival rates estimated by moving window based method show poor long-term survival for advanced stages of colorectal cancer (see Table 3.2), especially in older patients, that is consistent with previously published population estimates for the Czech Republic [21]. This phenomenon can be seen also in other European countries [38, 96] and still introduces a great challenge for new treatment modalities. This problem is even more pressing in the Czech Republic, as a large proportion of colorectal carcinomas are primarily detected in metastatic form [21].

There is no comparative evidence on stage-specific colorectal cancer recurrence rates based on the analysis of population-based data. The only data on cancer recurrence rates have been published for all stages combined [16, 36], and, in addition, without the consideration of the phase of cancer treatment. The rates associated with the risk of cancer recurrence in time proposed in this paper for Czech colorectal cancer patients, both non-terminal and terminal, show reasonable estimates validated by the leading clinicians of the Czech Society for Oncology. However, future verification on population-based or hospital-based level would be of a great value.

The total number of patients potentially requiring active anti-tumour therapy in the Czech Republic in 2011 represents almost 9,000 patients of which approximately one third is predicted for the most compelling treatment of the generalised disease. Considering the anticipated impact on financial budget associated with the predicted numbers, it is obvious that the principal strategy of the fight against colorectal cancer in the Czech Republic should focus predominantly on the prevention of colorectal cancer, which may be accomplished via prevention programmes, organised screening programmes, and further improvements in diagnostic methods.

Five-Year Survival Rates of Cancer Patients in the Czech Republic

4

This chapter aims to document the applicability of population-based methods for survival assessment on the data of the Czech National Cancer Registry. The chapter also presents the survival benchmarks for Czech cancer patients which can be readily used for health care assessment, and provides an overview of the survival rates achieved since 1990. In addition, a comparison is provided between the Czech and European data concerning the five-year relative survival rates in selected cancer diagnoses.

The survival analysis on population level is not simple from the methodical point of view, and is still the subject of ongoing research. For this reason, the obtained results must be assessed very carefully, as survival rates represent a parameter that integrates many factors; and these factors, at the time of death, might not be necessarily related to the treatment of the original malignant tumour. Survival rates are indicators of very complex population relations and trends, and improvements in patient survival do not necessarily result from a more effective treatment. In the short-term, this might be the consequence of better diagnostic methods which make it possible to detect less advanced stages and to achieve better treatment results [98, 25].

4.1 Reference data set and the time period for assessing population-based survival

The CNCR database was introduced in chapter 1 as a solid background for the assessment of achieved survival rates. The Czech population-based data have been processed in two ways, each of them having different interpretation:

- **Analyses involving all patients with non-zero survival**. This data set also involves patients who have not undergone treatment for various objective reasons (early death, treatment contraindication, patient's refusal to treatment, etc.). Only the DCO records and records of tumours found at autopsy are dropped from the entire population-based registry. Those records with

zero survival are also dropped in international studies [6, 18], as not being relevant for survival analyses. The analysis of such a widely defined data set provides a representative epidemiological picture of cancer survival in the Czech Republic; however, this result has only limited information value for the analysis of cancer care results. Moreover, the clinical stage is not determined for an objective reason in many untreated patients, which means that the stratification of survival with respect to the disease stage is impossible.

- **Analyses based on data of treated patients only**. This data set includes all records bearing witness to health care results in facilities involved in the care of cancer patients. Some records are dropped, such as the records on patients who died early and their treatment was not started, as well as records on patients who have not been treated for other objective and known reasons (refusal to treatment, treatment not applied due to very high age or poor health condition, etc.). The analysis performed in this manner shows the survival experience of really treated patients and can be considered to be a benchmark for health care quality and the results. The principles setting the definition of such set of population data can be summarized as follows:
  - Selection of treated patients with complete (verified) diagnosis and clinical stage.
  - Records involved in the analysis must have non-zero survival time.
  - Survival is calculated separately for individual clinical stages; this implies that record on clinical stage is obligatory in this reference data set.
  - The analyses are performed on the most recent period for which the population-based data are available (in this case, 1995-2005).

All survival data presented in this chapter were acquired from the records of patients older than 14 years. From the perspective of this chapter, childhood tumours cannot be regarded as compatible with adulthood tumours; for this reason, the reference data set was reduced in accordance with the international studies EUROCARE-4 and CONCORD. Figure 4.1 shows the division of all population-based data from the Czech National Cancer Registry (CNCR). It is obvious that data in three layers have been processed, each of them allowing analysis with different interpretation value:

(i) **The data set A** can be considered to be epidemiologically representative of the Czech population of cancer patients. Survival rates calculated in this

group of patients correspond to the overall situation within the population, but without link to the cancer care results of health care facilities, as a relatively large proportion of records apply to *de facto* untreated patients.

(ii) **The data set B** meets the condition of complete diagnosis of malignant neoplasm; it includes both treated as well as untreated patients for objective reasons (treatment contraindication, very high age, refusal to treatment, extremely advanced disease, etc.). The survival achieved in this group of patients provides a realistic picture of the overall management of malignant tumours, particularly the ability of the health care system to ensure early diagnosis of cancer.

(iii) **The data set C** includes only treated patients and thus allows us to estimate reference survival rates which are actually achieved in patients treated with anticancer therapy. These estimates give evidence on the health care results in specific cancer care facilities, and can be actually used as benchmarks.

Figure 4.1 also points out that the Czech population-based data cannot be used for the survival assessment in some diagnoses (e.g. haemato-oncology diagnoses and malignant tumours of the central nervous system, CNS). This is due to insufficient diagnostic identification of these cancer groups in the minimal CNCR record. As for CNS tumours, information on their grades is not available, whereas in leukemias, the acute and chronic types cannot be credibly stratified. The only solution for these diagnoses is to collect more detailed records in separate clinical registries.

Figure 4.1 documents that the Czech population-based data are robust enough even for a detailed survival analysis, which is further confirmed by the data provided in Table 4.1. It is obvious that the numbers of patients in individual diagnostic groups are sufficient, as they are available in the order of thousands. There are also some less frequent diagnoses (in the order of hundreds), such as malignant neoplasms (MN) of the oesophagus (C15), MN of the liver and intrahepatic bile ducts (C22), MN of the pancreas (C25) and MN of the vulva and vagina (C51-C52). The data set size is also acceptable after stratification into clinical stages, hundreds to thousands of records being available in most diagnostic groups. Categories with less than 100 records are rare, such as MN of the pharynx (C09-C14) in clinical stage I, MN of the oesophagus (C15) in clinical stage I, MN of the liver and intrahepatic bile ducts (C22) in clinical stages I, II and III, MN of the pancreas (C25) in clinical stage II, and MN of the vulva and vagina (C51-C52) in clinical stage IV.

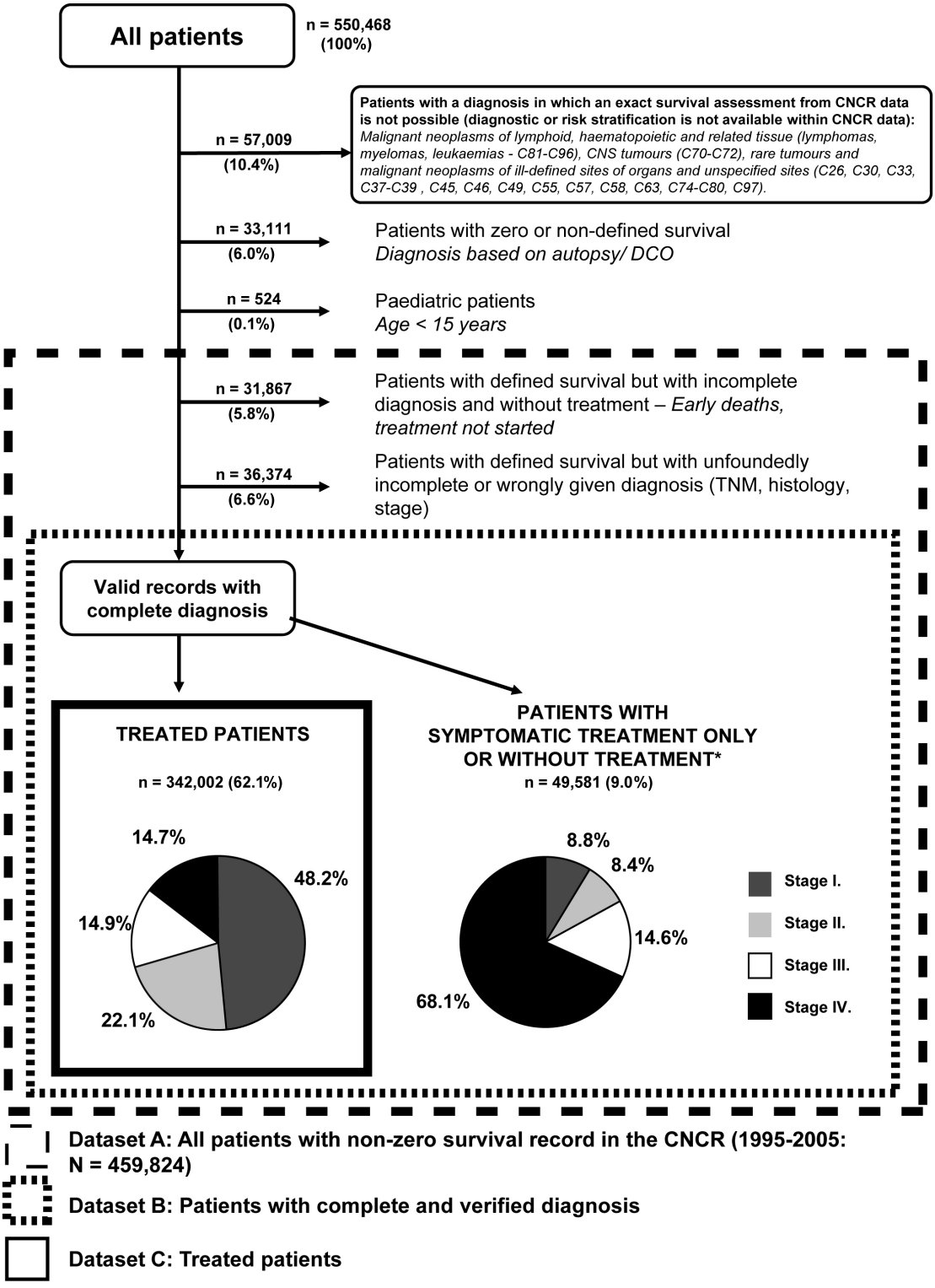

Figure 4.1: Definition of reference data set from the CNCR for the purpose of population-based survival analyses (1995 2005).

Table 4.1: Number of patients in data set C (treated patients), as defined in the CNCR; these data have been used in the calculation of five-year survival rates using 2003-2005 period analysis.

Diagnosis group	Data set C: Treated patients				
	Treated patients in total	Treated patients in stage I	Treated patients in stage II	Treated patients in stage III	Treated patients in stage IV
Oral cavity	2650	925	484	415	826
Pharynx	1853	114	200	423	1116
Oesophagus	824	63	266	268	227
Stomach	3588	1209	780	728	871
Colon and rectum	30623	7634	10993	7247	4749
Liver and intrahepatic bile ducts	388	28	48	106	206
Gallbladder and biliary tract	1250	289	325	169	467
Pancreas	1496	171	198	265	862
Larynx	2393	812	427	518	636
Bronchus and lung	12034	2046	1286	4143	4559
Melanoma of skin	8570	5622	1766	901	281
Breast	30012	10954	13605	3625	1828
Vulva and vagina	880	416	238	158	68
Cervix uteri	5446	3242	904	1023	277
Corpus uteri	8365	6714	753	664	234
Ovary and other uterine organs	4772	1865	404	1489	1014
Prostate	12216	2301	5616	1789	2510
Testis	2594	1734	509	351	*
Kidney and other organs of urinary tract	10252	4141	3208	1632	1271
Bladder	8354	5783	1657	467	447
Thyroid gland	3204	1700	936	355	213
Total	151764	57763	44603	26736	22662

* The new TNM classification does not define stage IV for diagnosis C62

Apart from the data set size, the age structure of cancer patient population is also important for reference survival outcomes, as the weights for the calculation of age-standardized survival rate estimate are derived from the age structure. The Czech data have been age-standardized in accordance with the method applied to international studies EUROCARE-4 and CONCORD (age groups: 15-44 , 45-54, 55-64, 65-74, and 75+ years). The weights have been calculated separately for individual diagnoses according to the age structure of Czech cancer patients.

It must be taken into account, however, that weights calculated on the Czech population are only meaningful when comparing the survival within the Czech Republic, and are not suitable for comparison with other populations. For the purpose of international analyses, weights common to the compared data sets must be defined; weights defined for the population of European cancer patients [19] might serve as an example.

4.2 Survival benchmarks for Czech cancer patients

The five-year survival rates of Czech cancer patients are assessed on the basis of comprehensively representative population-based data, this section presents and comments on the results acquired by analysis of the so-called C data set (treated patients), as defined in Figure 4.1 and in Table 4.1. The outcomes of the five-year relative and observed survival in treated Czech cancer patients are provided in Table 4.2, the five-year relative survival rates also shown in Figure 4.2. The relative survival rates, which are presented in Table 4.2, are further supplemented with survival estimates corresponding to individual stages (Table 4.3). The relative survival rates clearly separate diagnoses with generally better prognosis, such as the malignant neoplasm (MN) of thyroid gland (C73), MN of testis (C62), MN of the breast (C50) and MN of the skin (C43), from others. On the contrary, diagnoses with lower probability of long-term survival include malignant neoplasms of the digestive tract, such as MN of the pancreas (C25), MN of the liver and intrahepatic bile ducts (C22), MN of the oesophagus (C15), MN of the gallbladder and biliary tract (C23-C24), as well as MN of the bronchus and lung (C34). However, it is necessary to give several comments in order to avoid possible misinterpretation of these survival rates:

- All values of five-year survival rates, which are presented without stratification to clinical stages, cannot be simply compared among individual diagnoses, as they are significantly influenced by the proportional representation of clinical stages. Examples include the bladder carcinoma (C67) and prostate carcinoma (C61) which have reached almost the same values of

five-year relative survival (Table 4.2). However, this fact is explained in Table 4.3 and Figure 4.2 showing that better survival rates are achieved in all clinical stages in prostate carcinoma than in bladder carcinoma, if the clinical stages are assessed separately. The coincidence of relative survival rates without stage stratification in these different types of cancer (i.e., survival rates for prostate cancer appear to be as low as survival rates for bladder cancer) has been due to the significantly higher proportional representation of stages III and IV in prostate carcinoma (Table 4.1).

- Each point estimate of the five-year survival rate is supplemented with a 95% confidence interval. The variability of the estimate is affected by the number of patients and, obviously, by the heterogeneity of survival within the assessed group. In the Czech population-based data, a relatively high variability in survival estimates has only been observed in less frequent neoplasms, such as the MN of vulva and vagina (C51-C52), and MN of the liver and intrahepatic bile ducts (C22).

- In some less serious diagnoses which are detected at an early stage, the relative survival rate can approach or even exceed 100 %, for example as a result of more intensive medical care, self-control, and a more healthy lifestyle in these patients. In this case, the five-year observed survival becomes important, providing information on the overall mortality of these patients.

The comparison between relative and observed survival rates provides indirect information on the age structure of patients within a given diagnosis group. In diagnoses typical of younger age groups (such as MN of testis), both values are almost equal. On the contrary, in diagnoses typical of older patients (e.g. MN of prostate), the relative survival is visibly higher than the observed survival, which can be easily explained: the relative survival is related to the survival probability in a generally older population.

4.3 Survival rates achieved in all Czech cancer patients

The previous chapter presented survival estimates calculated solely on a reference set of patients, marked as data set C in Figure 4.1, i.e. from the records of treated cancer patients. These calculations have been performed as a result of an effort to provide such reference survival rates which could be used in the assessment of health care results in clinical practice. The calculation of five-year survival rate, however, is not limited to treated patients only; routine population-based estimates take into account a more widely defined group of patients. Survival estimates, frequently without stratification into clinical stages, are typical

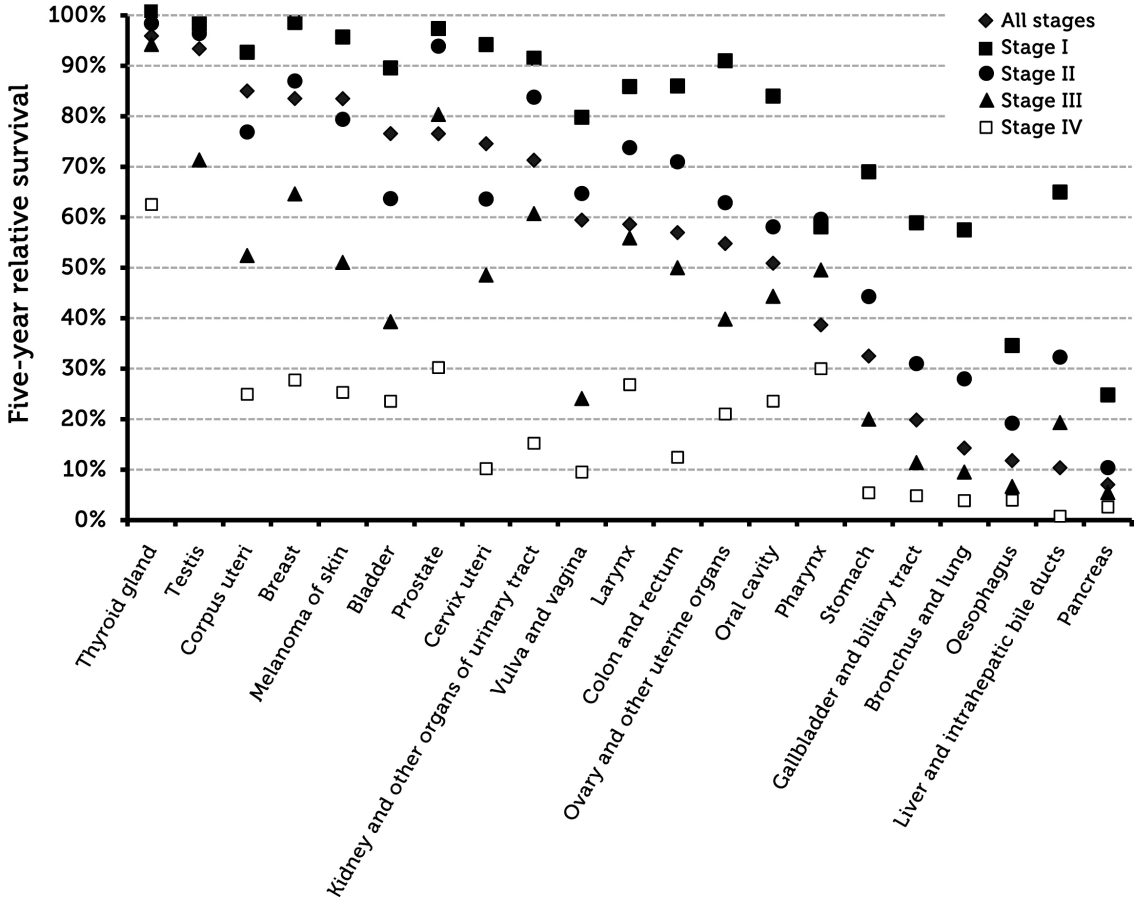

Figure 4.2: Comparison of five-year relative survival rates in treated Czech cancer patients diagnosed in different clinical stages (period analysis: 2003-2005).

for international studies, as the contents and quality of cancer registries in many countries do not allow the conductance of detailed analyses. The population-based estimates which take into account *de facto* all patients included in CNCR who have non-zero survival has been also prepared (Figure 4.1; data set A).

Table 4.4 provides a comparison of the five-year relative survival rates of cancer patients calculated on three different data sets. Logically, the largest difference in the resulting survival rates can be anticipated when comparing data set A with data set C. The difference is particularly significant in diagnoses where there is a problem with early diagnosis of less advanced clinical stages. The comparison between survival rates in all patients and in treated patients shows that worse survival of the former group can be attributed to the following factors:

(i) **The proportion of patients diagnosed in clinical stage IV and, generally, the proportion of patients with very advanced disease, resulting in non-treatment and early death.** This influence is particularly significant in ma-

lignant neoplasms of the digestive tract, and leads to lower survival rates observed over the whole population of patients, if compared to survival rates in patients diagnosed in less advanced stages. On the contrary, there is a low percentage of patients with untreated malignant melanoma (MN) of the skin (C43), MN of the breast (C50), MN of testis (C62), and MN of thyroid gland (C73).

(ii) **The proportion of records with wrongly recorded or unfoundedly incomplete diagnosis of malignant tumour in the CNCR.** These records have faults pursuant to errors in the population-based registry and cannot be included in stratified analysis according to the clinical stage, or they even challenge the cancer diagnosis. These records are, therefore, also omitted from calculations resulting in clinical reference survival rates. The proportion of faulty records does not correlate with the seriousness of the disease; in particular, higher error rate can be observed for prostate carcinoma (C61) and bladder carcinoma (C67).

4.4 Time trends in population-based survival of Czech cancer patients

The assessment of time trends in population-based survival brings forward plenty of useful information, although it is rather demanding from the methodical point of view. Its importance is obvious: the five-year relative survival is a benchmark of the health care results and its development over time provides valuable information [69, 23, 25]. Its interpretation is, however, complicated by a number of various factors over time, such as prevention programmes, development in diagnostic methods or the anti-cancer therapy itself [25, 98].

Similar to the comparison of different population of patients, survival rates in different periods of time must be also age-standardized [19, 54]. Changes in the age structure of patients can occur due to many factors, demographic developments or organized screening programmes playing the most important role. The impact of screening on five-year relative survival rates has been reported many times, e.g. in [25, 6, 95], particularly in connection with the so-called "overdiagnosis bias". Overdiagnosis refers to cases where a cancer screening programme reveals a tumour (often with a very good prognosis) which, however, would not show any symptoms during the patient's life, because the patient would die earlier due to another condition. These overdiagnosed tumours markedly improve the survival assessment, even though there is no benefit to the patients themselves. The overdiagnosis bias is an extreme case of the so-called "length bias"

which is partiality stemming from the fact that screening programmes primarily detect slowly-growing tumours, which have better prognosis. Detection of cancer by screening programmes also introduces the so-called "lead time bias", as screening programmes, by definition, detect asymptomatic tumours, shifting the time of diagnosis. In other words, improvements in survival rates do not necessarily mean that an early diagnosis has a beneficial effect on the disease fatality, and does not have to reflect improvements in health care at all. These problems with interpretation can be avoided if survival assessments are done for treated patients only, and separately for different clinical stages.

The Czech population-based data contain sufficiently long time series of the Czech National Cancer Registry (CNCR, standardized data collection since 1977), making it possible to compare the survival rates over time. This chapter provides an example of such analyses, performed solely on clinically relevant records of patients in which the diagnosis of malignant tumour was completed and verified, and who were treated. The five-year relative survival rates are assessed on sets of patients defined as follows:

- 2003-2005 period analysis - reference survival rates (or benchmarks)
- 2000-2002 period analysis
- Cohort analysis of patients diagnosed in 1995-1999
- Cohort analysis of patients diagnosed in 1990-1994

In order to be comparable, the calculated values are age-standardized in relation to the age structure of cancer patients corresponding to the 2003-2005 period.

Table 4.5 summarizes the development of five-year relative survival rates over time, diagnoses being sorted according to the importance of change in survival rates over time in the periods 1990-1994 and 1995-1999. The most important shift has occurred after the period 1990-1994, although improvements have been recorded in a number of diagnostic groups also after year 2000. However, the changes in five-year relative survival rates over time vary considerably among diagnoses. Significant improvements have been observed in prostate carcinoma (C61), breast carcinoma (C50) and skin melanoma (C43). Unfortunately, there are several diagnoses which show relatively high mortality and in which changes in survival rates are negligible, such as the carcinoma of pancreas (C25), carcinoma of stomach (C16), malignant neoplasm of gallbladder and biliary tract (C23-C24) or lung cancer (C34).

Obviously, the interpretation of data in Table 4.5 is limited by all factors mentioned above, namely the ongoing screening programmes, improvements in diagnostic methods, as well as modifications in TNM classification. For this reason,

comparison of the survival rates over time has been supplemented with analysis for individual clinical stages (Figure 4.3). This analysis also confirms improvement in five-year relative survival rates, particularly in malignant neoplasms diagnosed with clinical stages I and II. These data unambiguously document a positive development in the Czech cancer care over the last 10-15 years. There have also been improvements in survival rates for clinical stages III, with the exception of lung cancer (C34) and MN of the kidney and other structures of the urinary tract (C64-C66, C68).

As regards malignant tumours diagnosed primarily at clinical stage IV, visible improvements in treatment results are very rare over 10-15 years. The only statistically significant improvement has been observed in breast cancer (C50) which correlates with rapid development of treatment modalities in metastatic breast cancer after the year 2000. Results in other diagnoses have confirmed that the treatment of tumours in clinical stage IV continues to present a formidable challenge. This problem is even more pressing in the Czech Republic, as a large proportion of malignant tumours is primarily detected in metastatic form (Table 4.1). The principal strategy of the fight against cancer in the Czech Republic, therefore, consists in the prevention of advanced stages of malignant neoplasms, which may be accomplished via prevention programmes, organized screening programmes, and further improvements in diagnostic methods.

4.5 Comparison of five-year relative survival rates of Czech cancer patients with international data

A number of international projects have been making efforts to interconnect and centralize the European data; several international organisations have been established with the aim to assess the health care quality and performance, and a number of journals dealing with this issue have surfaced. It is not about the so-called "scientific marketing" of own results; these comparisons provide very important methodical and clinical conclusions which contribute to improvements in health care in each individual country.

The Czech Republic cannot stay out of the growing international cooperation, and there is no reason for it either, as the Czech oncology is very well equipped with population-based data, and has adequate methodical background in data collection and assessment. It is very important, however, to pay attention to correct interpretation of the data and the results being compared. In the case of survival data on Czech cancer patients, estimates of the relative survival rates conducted on all cancer patients (with a non-zero survival in CNRC) can be readily used for the purpose of international comparison. This is due to the fact that

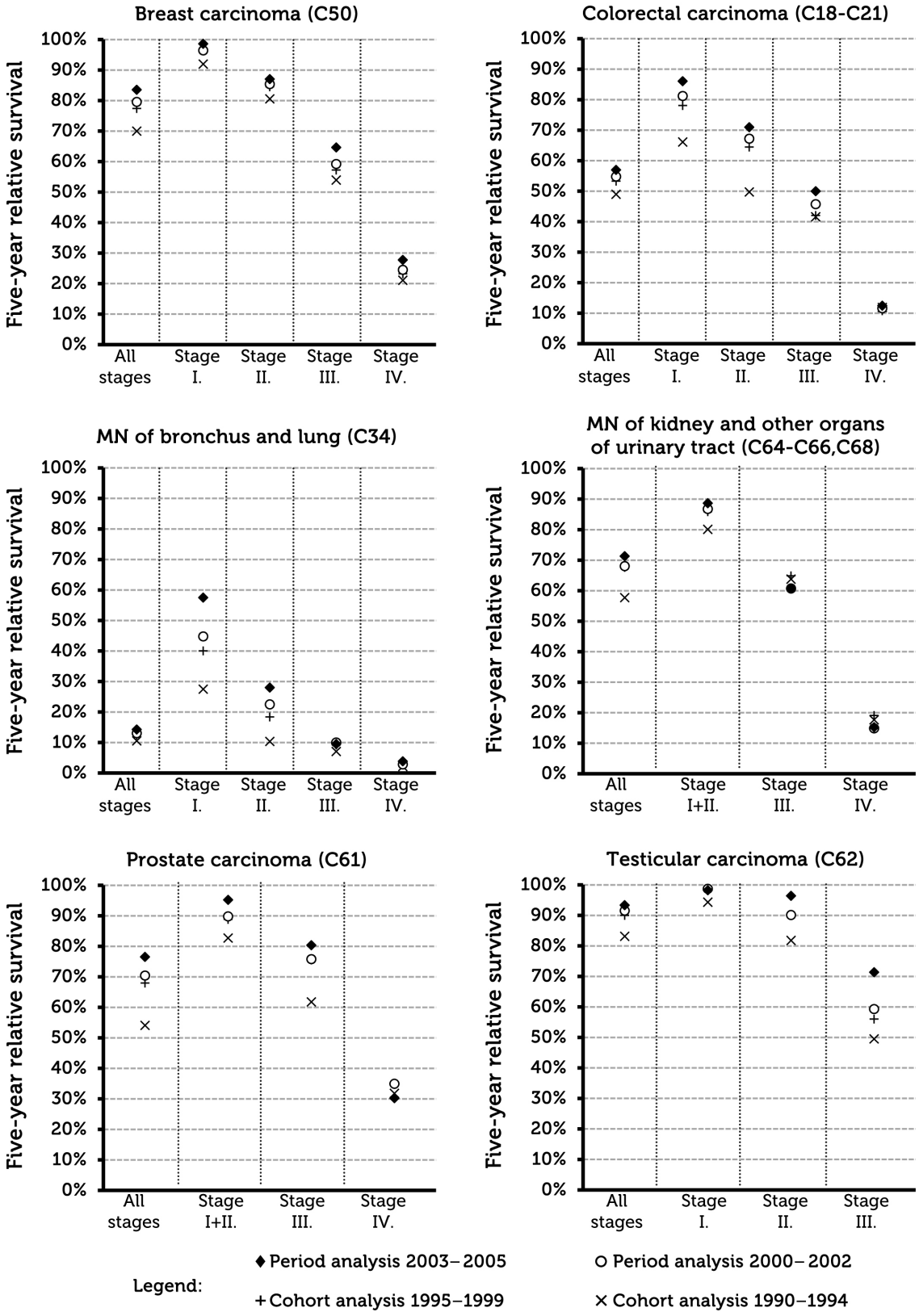

Figure 4.3: Comparison of five-year relative survival rates in treated cancer patients - selected diagnoses sorted by clinical stages.

international data do not provide population-based comparisons which would respect the clinical stage of malignant tumours, nor do they provide assessment focused on treated patients. Population-based data containing such details are not even available in many developed countries. For this reason, a comparison of the population-based survival rates is most frequently encountered, which renders a picture of the overall epidemiological situation, but fails to mention anything about the cancer care results in specific health care facilities.

The Czech data can be surely compared with data from the EUROCARE-4 study which is excellent from the statistical point of view. More specifically, the Czech data might be compared with the main part of the EUROCARE-4 study, containing survival analysis of patient cohort diagnosed in 1995-1999, and with the part focused on 2000-2002 period analysis. Furthermore, the Czech results can be compared with the outcomes of studies published on the population-based data of individual European countries or regions, or even globally. In brief, international data selected for the assessment of survival rates in the Czech Republic can be summarized as follows:

(i) EUROCARE is an international activity connecting data from population-based cancer registries of selected European countries, with the aim to get an overview of the achieved survival rates. Considering the diversity of approaches in data registration in various countries, many of which do not fully comply with representative registries, this is a very demanding task. This has also been confirmed by the fact that the final assessment of the EUROCARE-4 study [6] does not involve all European countries (data from 23 countries are presented, including 83 population-based cancer registries). The EUROCARE-4 study deals with the comparison of survival rates calculated using the cohort analysis over 1995-1999, and outputs calculated using the 2000-2002 period analysis are also available [95].

(ii) Apart from the EUROCARE-4 study, there are other recent studies dealing with the presentation of relative survival rates in individual countries across Europe and worldwide, see for example [89, 12, 38, 55, 18, 79]. However, not all such data can be actually used for comparison with five-year relative survival rates in Czech cancer patients, as these studies differ in methodology or the assessed time periods.

Figures 4.4 and 4.5 and Table 4.6 present the comparison of five-year relative survival rates estimated for Czech cancer patients with the results published in the EUROCARE-4 study. Figure 4.4 is dedicated to comparison of the five-year relative survival rates based on patient cohort diagnosed in 1995-1999, i.e. comparison with the main part of the EUROCARE-4 study, whereas Figure 4.5

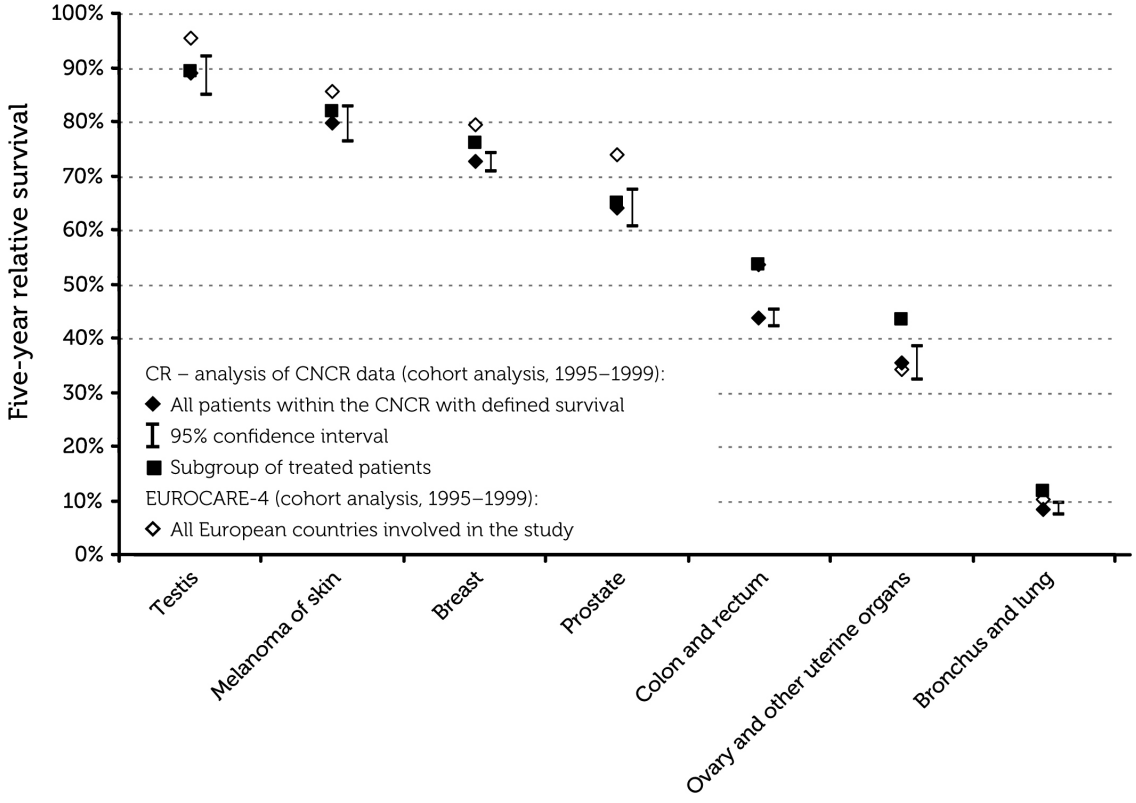

Figure 4.4: Comparison of five-year relative survival rates in Czech cancer patients with rates published in the EUROCARE-4 study (cohort analysis of patients diagnosed in 1995-1999).

compares the five-year relative survival rates estimated using the 2000-2002 period analysis. As regards the 1995-1999 cohort analysis results, the pooled data, collected by the European cancer registries, show higher five-year relative survival rates than the data of all patients within CNCR with defined survival for all presented diagnoses except for malignant neoplasms (MN) of the ovary (C56) and MN of the bronchi and lungs (C34). Considering the more recent results represented by the 2000-2002 period analysis, similar differences can be observed in the five-year relative survival rates between the pooled European data and CNCR data of all patients with defined survival for diagnoses analysed in both parts of the EUROCARE-4 study. On the other hand, consistency in the five-year relative survival rates between the pooled European data and CNCR data can be seen in the diagnoses included only in the EUROCARE-4 period analysis (see Figure 4.5 and Table 4.6), namely for MN of cervix uteri (C53), MN of uterus (C54), MN of kidney and other structures of the urinary tract (C64-C66, C68), and MN of thyroid gland (C73).

There is no easy interpretation of the differences in population-based survival

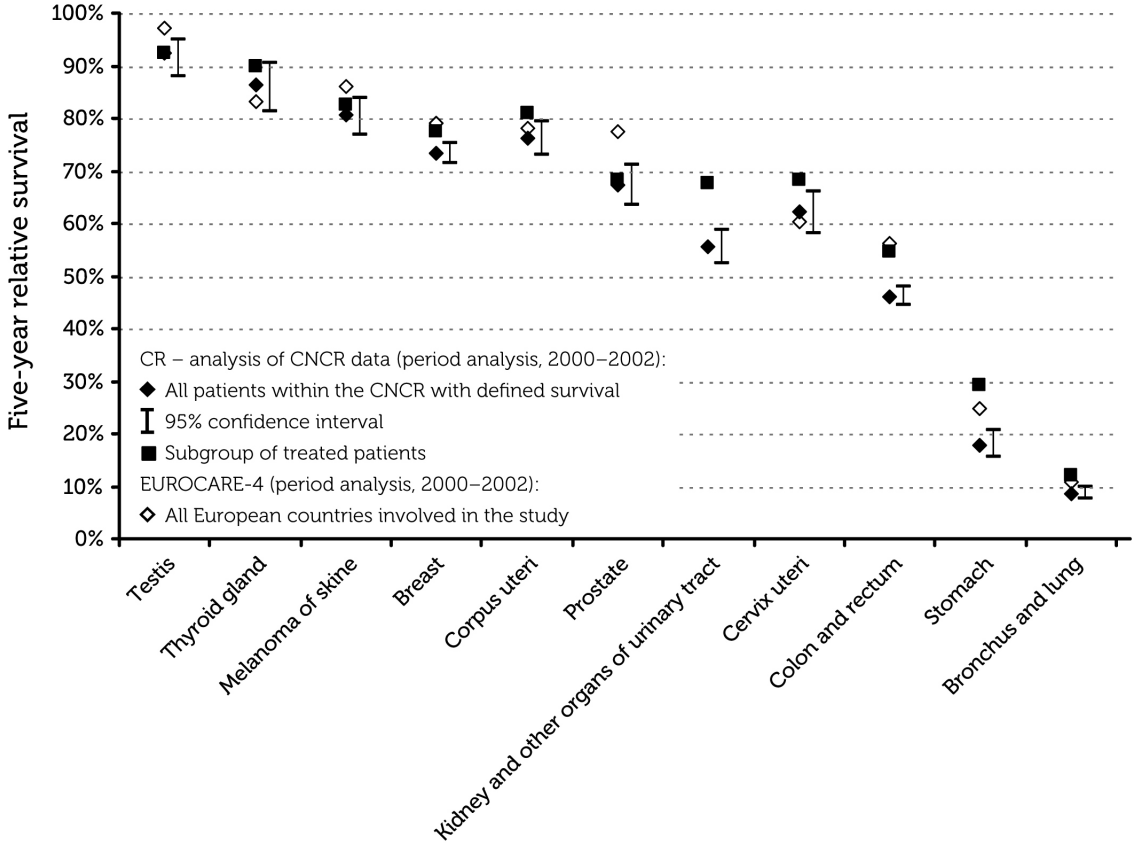

Figure 4.5: Comparison of five-year relative survival rates in Czech cancer patients with rates published in the EUROCARE-4 study (2000-2002 period analysis).

rates between two or more populations without proper stratification to possible confounding factors of which clinical stage and anticancer therapy are the best examples. It has to be stressed that the final results are often affected by demographic differences as well as differences in the age structure of patients, which can always be found between the populations being compared. The primary goal of the EUROCARE study is to make comparable data from registries which describe population with different demographic structures. This must be, therefore, reflected in the methodology of standardisation.

- The assessed period itself has tremendous impact on the calculation of survival rates. Fortunately, the available EUROCARE-4 data allows for comparison with the Czech data collected over the same periods. When interpreting the comparison of survival rates, one does have to bear in mind that there have been improvements in survival rates of the Czech cancer patients over time, and that the comparison of older data sets does not correspond to the more recent treatment results.

- The methodology of survival estimate is another factor impacting the final survival rates. The Czech Society for Oncology has been processing population-based data in the same manner as international studies, and the results are fully compatible from the methodical point of view.

- The applied age standardisation can also affect survival rates. Reference five-year relative survival rates suggested for the Czech Republic have been standardized with regard to proportional representation of individual age groups (15-44, 45-54, 55-64, 65-74 and 75-99 years), taking into account solely Czech patients. First of all, survival rate was calculated for each of the defined groups; and subsequently, a weighted average was calculated, corresponding to the structure of the Czech data. Authors of the EUROCARE-4 study used weights derived from the age structure of the population of European cancer patients form the EUROCARE-2 study [19]. This weighing, however, partially disadvantages populations with a different (particularly younger) age structure compared to the structure of European cancer patients. In many diagnoses, this also applies to the Czech population.

- The representation of clinical stages within the populations being compared is another factor which is frequently dismissed, but strongly affects survival rates. Outcomes that do not analyse survival rates with respect to clinical stages are merely a benchmark of the epidemiological situation in individual countries, and cannot serve as a benchmark of treatment results in cancer care facilities.

- The previous point is closely related to stratification of patients depending on whether they have or have not undergone anticancer treatment (and if not, then why?). If we want to compare the treatment results in individual countries in a consistent manner, we have to confine our assessments to those patients who have been admitted to a health care facility and who have been provided treatment.

To conclude this chapter, it is important to highlight that the differences between five-year relative survival rates in Czech patients and in the EUROCARE-4 study, which have been observed particularly in diagnoses such as MN of colon (C18-C21), MN of breast (C50), malignant melanoma of skin (C43), MN of prostate (C61), and MN of testis (62), cannot be unequivocally attributed to a poorer health care provided to Czech cancer patients. One possible explanation for this difference is a high proportion of patients diagnosed at stage IV, which is a consequence of a poor education of patients about early symptoms of the disease and the non-existence of organized screening programmes in the 1990s, rather than a poor quality of health care. Furthermore, a significant improvement

in survival rates has been observed in many diagnoses since the 1990s, and the comparison of retrospective values does not therefore correspond to the current situation in Czech cancer care. Last but not least, the assessment of health care results should be based on the comparison of survival rates in patients who have undergone anticancer treatment. Only results from a set of treated patients can actually bear witness to the quality of health care, as such set does not involve patients who have not been treated for objective reasons (treatment contraindication, very high age, patient's refusal to treatment, very advanced stage of disease), or patients with incomplete diagnosis.

Table 4.2: Five-year observed and relative survival rates in treated Czech cancer patients accompanied with 95% confidence intervals (2003-2005 period analysis).

		Observed survival rate		Relative survival rate	
Diagnosis group	N	Five–year survival	95% CI	Five–year survival	95% CI
Oral cavity	2,650	43.6	38.3-48.7	50.9	44.6-57.0
Pharynx	1,853	35.1	29.3-40.8	38.7	32.1-45.3
Oesophagus	824	10.3	6.3-15.7	11.8	7.1-18.2
Stomach	3,588	27.8	24.1-31.6	32.5	28.2-36.9
Colon and rectum	30,623	47.9	46.3-49.4	57.0	55.1-58.8
Liver and intrahepatic bile ducts	388	9.0	3.9-16.6	10.4	4.5-19.1
Gallbladder and biliary tract	1,250	16.9	12.5-21.9	19.9	14.6-25.8
Pancreas	1,496	6.6	4.0-9.9	7.0	4.3-10.6
Larynx	2,393	52.4	46.7-57.7	58.6	52.1-64.7
Bronchus and lung	12,034	12.7	11.4-14.1	14.3	12.8-15.9
Melanoma of skin	8,570	75.7	72.8-78.2	83.5	80.2-86.5
Breast	30,012	76.5	75.1-78.0	83.5	81.8-85.1
Vulva and vagina	880	51.7	41.9-59.7	59.4	48.2-69.0
Cervix uteri	5,446	72.6	69.3-75.5	74.6	71.1-77.7
Corpus uteri	8,365	78.3	75.8-80.5	85.0	82.2-87.5
Ovary and other uterine organs	4,772	52.2	48.3-55.9	54.8	50.6-58.8
Prostate	12,216	59.0	56.6-61.4	76.5	73.3-79.7
Testis*	2,594	92.1	89.6-93.9	93.4	90.8-95.2
Kidney and other organs of urinary tract	10,252	63.2	60.4-65.8	71.3	68.1-74.3
Bladder	8,354	63.7	60.8-66.4	76.6	72.9-79.9
Thyroid gland	3,204	91.8	88.6-94.0	95.9	92.3-98.5

* In malignant neoplasms of testis, the observed and relative survival rates were only estimated for patients in age groups not exceeding 65 years.

Table 4.3: Five-year relative survival rates in treated Czech cancer patients, according to diagnoses and clinical stage (2003-2005 period analysis).

	Stage I			Stage II			Stage III			Stage IV		
Diagnosis group	N	5-year RSR	95% CI	N	5-year RSR	95% CI	N	5-year RSR	95% CI	N	5-year RSR	95% CI
Oral cavity	925	84.0	72.1-93.3	484	58.1	42.3-71.7	415	44.4	29.9-58.4	826	23.6	16.3-32.1
Pharynx	114	58.1	26.8-84.2	200	59.6	36.4-76.4	423	49.6	35.2-62.6	1,116	30.0	22.4-38.0
Oesophagus	63	34.6	12.3-55.9	266	19.2	9.0-33.2	268	6.7	1.9-16.2	227	4.0	0.9-11.6
Stomach	1,209	69.0	58.9-77.4	780	44.3	33.7-54.5	728	20.0	13.2-28.0	871	5.4	2.3-10.9
Colon and rectum	7,634	86.0	82.4-89.3	10,993	71.0	67.8-74.0	7,247	50.0	46.1-53.9	4,749	12.5	10.1-15.2
Liver and intrahepatic bile ducts	28	65.0	38.6-87.4	48	32.3	13.2-54.1	106	19.3	10.6-34.9	206	0.8	0.1-3.9
Gallbladder and biliary tract	289	58.9	39.8-76.0	325	31.0	17.6-44.3	169	11.4	4.6-23.2	467	4.9	1.9-10.3
Pancreas	171	24.8	13.1-38.1	198	10.4	3.3-23.3	265	5.5	1.3-14.8	862	2.6	1.0-5.7
Larynx	812	85.9	73.1-94.5	427	73.8	60.2-85.1	518	55.9	43.1-67.9	636	26.8	18.9-36.2
Bronchus and lung	2,046	57.5	50.6-63.8	1,286	28.0	21.7-34.8	4,143	9.5	7.6-11.8	4,559	3.9	2.7-5.3
Melanoma of skin	5,622	95.7	92.3-98.4	1,766	79.4	71.0-86.6	901	51.1	40.4-61.3	281	25.3	10.6-44.3
Breast	10,954	98.6	96.5-100.4	13,605	87.0	84.6-89.3	3,625	64.6	59.0-69.9	1,828	27.7	21.8-34.1
Vulva and vagina	416	79.8	65.5-90.2	238	64.7	41.5-81.9	158	24.1	8.1-48.6	68	9.5	1.6-27.2
Cervix uteri	3,242	94.2	90.9-96.7	904	63.6	53.7-72.3	1,023	48.5	40.1-56.7	277	10.2	3.7-21.9
Corpus uteri	6,714	92.7	89.9-95.1	753	76.9	64.9-85.1	664	52.4	40.6-62.3	234	24.9	12.6-39.1
Ovary and other uterine organs	1,865	91.0	85.0-95.4	404	62.9	47.2-75.7	1,489	39.8	32.8-47.0	1,014	21.0	15.1-27.8
Prostate	2,301	97.4	91.0-102.8	5,616	93.9	88.8-98.4	1,789	80.4	72.0-87.8	2,510	30.2	25.4-35.4
Testis*	1,734	98.3	95.6-99.7	509	96.4	91.4-99.0	351	71.4	61.9-78.5	**	**	**
Kidney and other organs of urinary tract	4,141	91.5	86.2-95.9	3,208	83.8	78.3-88.4	1,632	60.7	52.4-68.3	1,271	15.3	10.7-20.7
Bladder	5,783	89.6	85.4-93.2	1,657	63.7	55.3-71.0	467	39.3	27.1-51.4	447	23.5	12.9-36.3
Thyroid gland	1,700	100.1	95.3-102.2	936	98.4	90.9-102.0	355	94.2	82.9-100.8	213	62.6	42.4-77.2

* In malignant neoplasms of testis (C62), the observed and relative survival rates were only estimated for patients in age groups not exceeding 65 years.

** The new TNM classification does not define stage IV for diagnosis C62

Table 4.4: Five-year relative survival rates calculated on three defined reference data sets (A, B, and C: see Figure 4.1); 2003-2005 period analysis.

Diagnosis group	Data set A: all patients recorded in CNCR with non-zero survival			Data set B: patients with complete and verified diagnosis			Data set C: treated patients with complete and verified diagnosis		
	N	5–year RSR	95% CI	N	5–year RSR	95% CI	N	5–year RSR	95% CI
Oral cavity	3,167	46.7	41.1-52.2	2,755	49.0	43.0-55.0	2,650	50.9	44.6-57.0
Pharynx	2,262	35.7	30.0-41.6	1,973	36.9	30.7-43.1	1,853	38.7	32.1-45.3
Oesophagus	1,464	9.4	6.1-13.8	1,071	9.7	6.0-14.8	824	11.8	7.1-18.2
Stomach	5,905	21.8	19.0-24.6	4,676	23.9	20.7-27.2	3,588	32.5	28.2-36.9
Colon and rectum	36,715	50.0	48.3-51.6	32,743	52.2	50.5-53.9	30,623	57.0	55.1-58.8
Liver and intrahepatic bile ducts	1,750	4.5	2.7-7.0	903	5.0	2.5-8.9	388	10.4	4.5-19.1
Gallbladder and biliary tract	2,826	10.6	8.1-13.4	1,931	12.8	9.6-16.5	1,250	19.9	14.6-25.8
Pancreas	4,313	4.3	3.0-5.9	2,909	4.7	3.0-6.8	1,496	7.0	4.3-10.6
Larynx	2,808	54.3	48.5-60.0	2,553	56.2	50.0-62.0	2,393	58.6	52.1-64.7
Bronchus and lung	18,742	10.5	9.5-11.6	15,741	11.2	10.1-12.5	12,034	14.3	12.8-15.9
Melanoma of skin	9,347	81.8	78.6-84.8	8,626	82.9	79.7-85.9	8,570	83.5	80.2-86.5
Breast	32,886	80.4	78.8-82.0	30,373	82.3	80.6-83.9	30,012	83.5	81.8-85.1
Vulva and vagina	1,152	55.6	46.0-64.0	926	56.4	45.6-65.8	880	59.4	48.2-69.0
Cervix uteri	6,525	70.5	67.3-73.5	5,673	72.7	69.4-75.9	5,446	74.6	71.1-77.7
Corpus uteri	10,439	81.6	79.0-84.1	8,570	83.5	80.6-86.0	8,365	85.0	82.2-87.5
Ovary and other uterine organs	5,940	49.4	45.7-52.9	5,048	51.5	47.6-55.3	4,772	54.8	50.6-58.8
Prostate	19,305	74.4	71.9-76.9	13,927	75.3	72.3-78.2	12,216	76.5	73.3-79.7
Testis*	2,877	93.5	91.1-95.3	2,597	93.2	90.6-95.1	2,594	93.4	90.8-95.2
Kidney and other organs of urinary tract	12,794	62.9	60.1-65.6	11,016	65.3	62.3-68.3	10,252	71.3	68.1-74.3
Bladder	11,267	72.8	69.7-75.7	8,600	74.8	71.2-78.1	8,354	76.6	72.9-79.9
Thyroid gland	3,736	94.6	91.2-97.1	3,239	94.8	91.3-97.4	3,204	95.9	92.3-98.5

* In malignant neoplasms of testis (C62), the observed and relative survival rates were only estimated for patients in age groups not exceeding 65 years.

Table 4.5: Changes in five-year relative survival rates in treated Czech cancer patients in the periods 1990-1994 and 1995-1999 and in the periods 2000-2002 and 2003-2005.

Diagnosis group	Cohort analysis 1990-1994		Cohort analysis 1995-1999		Δ	Period analysis 2000-2002		Period analysis 2003-2005		Δ
	5–year RSR	95% CI	5–year RSR	95% CI		5–year RSR	95% CI	5–year RSR	95% CI	
Prostate	53.0	48.8-57.2	67.9	64.6-71.1	14.9	70.5	66.8-74.1	76.5	73.3-79.7	6.0
Melanoma of skin	71.5	67.6-75.2	81.7	78.5-84.7	10.2	82.6	79.0-85.9	83.5	80.2-86.5	0.9
Kidney and other organs of urinary tract	57.7	54.1-61.3	67.8	64.8-70.6	10.1	68.0	64.6-71.2	71.3	68.1-74.3	3.3
Breast	70.1	68.3-72.0	77.6	76.0-79.2	7.5	79.6	77.8-81.4	83.5	81.8-85.1	3.9
Testis	83.0	79.2-86.2	90.0	87.2-92.3	7.0	91.6	88.5-93.8	93.4	90.8-95.2	1.8
Ovary and other uterine organs	46.6	42.6-50.6	53.0	49.3-56.7	6.4	53.4	49.2-57.4	54.8	50.6-58.8	1.4
Thyroid gland	87.3	82.0-91.6	93.4	89.7-96.3	6.1	94.2	90.7-96.9	95.9	92.3-98.5	1.7
Pharynx	26.6	20.9-32.8	32.6	26.8-38.7	6.0	36.0	29.1-43.2	38.7	32.1-45.3	2.7
Oesophagus	4.8	1.9-10.7	10.1	5.7-16.7	5.3	11.6	6.5-20.1	11.8	7.1-18.2	0.2
Bladder	68.7	63.4-73.7	74.0	70.4-77.5	5.3	75.1	71.0-78.9	76.6	72.9-79.9	1.5
Colon and rectum	49.0	47.0-51.0	53.3	51.5-55.0	4.3	54.8	52.8-56.7	57.0	55.1-58.8	2.2
Gallbladder and biliary tract	17.3	6.5-33.0	21.6	16.2-27.7	4.3	21.9	16.0-28.3	19.9	14.6-25.8	-2.0
Bronchus and lung	10.1	8.8-11.5	12.3	11.0-13.7	2.2	13.0	11.5-14.6	14.3	12.8-15.9	1.3
Cervix uteri	72.4	69.4-75.3	74.6	71.6-77.5	2.2	73.7	70.2-77.1	74.6	71.1-77.7	0.9
Corpus uteri	80.3	77.5-82.9	82.3	79.7-84.7	2.0	86.0	83.1-88.6	85.0	82.2-87.5	-1.0
Oral cavity	48.2	42.6-53.8	47.6	42.2-52.9	-0.6	52.1	45.5-58.6	50.9	44.6-57.0	-1.2
Stomach	30.6	26.7-34.6	29.9	26.2-33.7	-0.7	29.6	25.4-33.9	32.5	28.2-36.9	2.9
Pancreas	8.4	1.4-24.2	7.2	4.1-12.2	-1.2	6.7	3.9-11.0	7.0	4.3-10.6	0.3
Larynx	54.3	48.8-59.7	52.7	47.3-58.0	-1.6	54.9	48.3-61.2	58.6	52.1-64.7	3.7
Vulva and vagina	60.9	50.3-70.2	59.3	49.6-67.6	-1.6	62.5	51.6-71.4	59.4	48.2-69.0	-3.1
Liver and intrahepatic bile ducts	*	*	15.1	6.8-27.5	*	11.6	5.1-21.8	10.4	4.5-19.1	-1.2

* The table does not include five-year relative survival rates from the period 1990-1994 for MN of the liver and intrahepatic bile ducts (C22) diagnosis, as data from the period 1990-1994 contained less than 50 records of treated patients.

Table 4.6: Comparison of five-year relative survival rates in Czech cancer patients with survival rates published in the EUROCARE-4 study. Calculation was done by the cohort analysis of patients diagnosed in 1995-1999 and by the 2000-2002 period analysis. Five-year relative survival rates are age-standardized according to weights used in the EUROCARE-4 study [19].

Diagnosis group	Dataset A: All patients recorded within the CNCR with non-zero survival						Dataset C: Treated patients			
	Cohort analysis 1995-1999			Period analysis 2000-2002			1995-1999		2000-2002	
	Czech patients		Eurocare-4	Czech patients		Eurocare-4	Czech patients		Czech patients	
	5-year RSR	95% CI	5-year RSR	5-year RSR	95% CI	5-year RSR	5-year RSR	95% CI	5-year RSR	95% CI
Stomach	17.2	15.2-19.5	*	17.9	15.4-20.5	24.9	29.7	26.0-33.6	29.2	25.0-33.6
Colon and rectum	43.6	42.0-45.1	53.5	46.2	44.4-48.0	56.2	53.6	51.7-55.5	54.8	52.6-56.9
Bronchus and lung	8.5	7.5-9.5	10.2	8.7	7.6-10.0	10.9	11.6	9.9-13.5	12.1	10.2-14.3
Melanoma of skin	79.7	76.4-82.7	85.4	80.6	77.0-83.9	86.1	81.7	78.3-84.8	82.6	78.9-86.1
Breast	72.5	70.7-74.3	79.5	73.3	71.3-75.4	79.0	76.1	74.1-78.0	77.4	75.1-79.5
Cervix uteri	61.8	58.4-65.1	*	62.1	58.1-66.1	60.4	68.3	64.0-72.4	68.2	63.2-72.9
Corpus uteri	73.1	70.2-75.9	*	76.2	72.9-79.3	78.0	77.5	74.1-80.7	81.1	77.2-84.6
Ovary and other uterine organs	35.3	32.2-38.6	34.2	36.0	32.4-39.7	*	43.4	38.9-48.1	42.3	37.4-47.3
Prostate	64.0	60.5-67.4	73.9	67.4	63.4-71.3	77.5	64.9	60.3-69.6	68.3	63.0-73.4
Testis	88.8	84.8-91.9	95.5	92.3	88.0-95.0	97.3	89.3	84.9-92.5	92.3	87.6-95.0
Kidney and other organs of urinary tract	55.1	52.3-57.8	*	55.6	52.4-58.7	55.7	67.3	63.7-70.9	67.6	63.4-71.6
Thyroid gland	84.9	80.1-89.1	*	86.2	81.2-90.5	83.2	90.0	84.6-94.5	89.9	84.5-94.6

* The EUROCARE-4 study does not provide age-standardized five-year survival rates for these diagnoses [6, 95].

Regression model for cytogenetic or molecular response in patients with chronic myeloid leukemia

5

The aim of this chapter is to present a Cox regression model for the achievement of the complete cytogenetic or molecular response to a modern targeted therapy in patients in chronic phase of chronic myeloid leukemia (CML). The model is based on data coming from a population study involving approximately half of Czech and all Slovak CML patients treated since 2000.

5.1 Definition of the primary objective

Chronic myeloid leukemia is one of the myeloproliferative diseases, a clonal disorder characterized by a distinctive cytogenetic abnormality, the Philadelphia chromosome (Ph1). This abnormality, associated with the so-called *BCR-ABL* fusion gene is present in more than 95 % of CML patients [37]. Therapy of CML was more or less unsatisfactory for a long time as the only patient's option for a long term survival was to undergo a stem cell transplantation which is, on the other hand, also accompanied with high mortality from various reasons. However, the treatment of CML underwent a breakthrough in 1998 as the *BCR-ABL* tyrosine kinase inhibitor imatinib was introduced [2]. Although imatinib clearly provides a higher likelihood of achieving complete cytogenetic remissions, i.e. achieving of an undetectable form of the disease, than any other regular therapy, it cannot be regarded as a curative treatment.

The magnitude of reduction in CML burden is a key prognostic indicator for patients treated for CML with imatinib. It was shown in [77] that delayed achievement of cytogenetic and molecular response is associated with increased risk of progression among patients in chronic phase CML treated with imatinib. Thus, the objective of this study was to identify characteristics of CML patients associated with prolonged time to complete cytogenetic response (CCgR) or major molecular response (MMR) to imatinib therapy, which could further indicate the increased risk of disease progression. Strictly speaking, the time from start of the imatinib therapy to development of the CCgR or the MMR was chosen as the primary endpoint of this study. Patients, who left the imatinib therapy

due to death, stem cell transplantation, intolerance or due to the development of resistance to imatinib, were censored.

5.2 Data

As already mentioned in Chapter 4, the Czech population-based data cannot be used for the survival assessment in hemato-oncology diagnoses due to insufficient diagnostic identification of these cancer groups in the minimal CNCR record. As for leukemias, the acute and chronic types cannot be credibly stratified. Therefore, the data analysed in this chapter come from an international registry Camelia that has been founded by the Czech and the Slovak Societies of Hematology in 2004.

5.2.1 Camelia project

Camelia is an international, multicentre clinical registry for monitoring of incidence, treatment and treatment response in patients with CML in the Czech Republic and Slovakia. This non-interventional observational registry is organized by the Leukemia Section of the Czech Hematology Society, and provides the background for a study coordinated by the Masaryk University. At present, 10 Czech and Slovak clinical centres and one non-clinical centre are involved. The project aims to establish a highly representative database containing valuable data for clinical research, and to contribute to a better awareness of achieved treatment results. Analyses of epidemiological and clinical data allow the results of various therapeutic approaches to be regularly discussed; based on these discussion, centres may cooperate to develop generally applicable therapy guidelines.

In addition to basic characteristics identifying the patient and his disease the patient's record includes variables corresponding to disease development and individual phases of medical care. In general, patient record consists of the following components:

- Patient identification, CML confirmation examination
- Input clinical characteristics, Sokal and Hasford prognostic scores
- Blood count, blood differential, bone marrow examination
- Cytogenetics, clonal chromosome abnormalities
- Treatment modalities and their side-effects
- Dosing, response to therapy, disease progression

- Bone marrow stem cell transplantation.

Several variables including treatment dosing, concomitant treatment, development of clonal chromosome abnormalities, and side-effects can be recorded repeatedly during the follow-up period whereas the rest of the parameters can be recorded only once, at the time of diagnosis.

5.2.2 Patients included in the analysis

In total, 658 CML patients diagnosed in years 2000–2008 with extended data record were entered into the database. However, the set of 658 patients represents a mixture of patients treated with various treatment modalities in all phases of CML defined above. Therefore, only 330 CML patients treated with first-line imatinib were primarily considered for the analysis. The term first-line refers to the fact that these patients were given imatinib as the initial treatment, meaning that they had no regular CML therapy prior to imatinib. This step ensures a certain degree of consistency among the patients for the treatment response can be easily attributable to the effect of imatinib and not to any therapy administered before imatinib. Another advantage of selecting patients with imatinib in the first-line treatment is in the usability of characteristics measured at the time of diagnosis. Their effect can be attributable to the start of the imatinib therapy as the date of diagnosis is close to the date of initiation of imatinib therapy in the first-line patients. However, two more filters should be adopted to select the final set of patients:

(i) **Filter for the length of follow-up**. Patients with follow-up less than 12 months from the start of imatinib therapy were not considered in the analysis. The reason is the equal chance of achieving the CCgR or the MMR to imatinib therapy for all patients, i.e., according to the European LeukemiaNet guidelines [3], the patients should have been followed-up for at least 12 months from the start of imatinib therapy to make it possible for the CCgR or the MMR to occur.

(ii) **Filter for the key data missing**. Patients with missing values of key characteristics were not considered in the analysis. The key characteristics were defined as follows: date of birth, sex, date of diagnosis, date of initiation of imatinib therapy, Sokal and Hasford prognostic scores, blood count (used for definition of anemic patients), and imatinib dosing at the treatment start.

Moreover, a pilot analysis has revealed a strong association between the time to the CCgR or the MMR and the period of diagnosis represented by two time intervals 2000–2004 and 2005–2008, respectively, which is, however, very unlikely

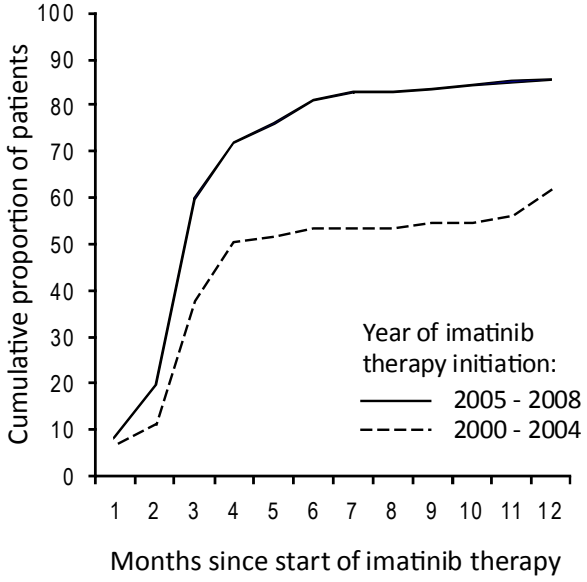

Figure 5.1: Cumulative proportion of patients with follow-up examination in first 12 months after imatinib therapy initiation.

to be true from the clinical perspective. As these two time periods correspond with the retrospective part (time period 2000–2004) and the prospective part (time period 2005–2008) of the Camelia project, respectively, it is more likely that this phenomenon is much more related to frequency and especially availability of clinical and cytogenetic follow-up examinations during the first 12 months after the start of the imatinib therapy, which is higher in the latter time period as can be seen in Figure 5.1. This explanation is also supported with the fact that, in total, the overall proportion of patients, who achieved the CCgR or the MMR, is similar in time periods 2000–2004 (73.2 %) and 2005–2008 (77.6 %). The difference in the quality of patient follow-up between the two time periods has led to selection of the latter time period (2005–2008) for the modelling. An overall scheme summarizing the selection of patients is given in Figure 5.2.

Finally, the set of N=197 patients was considered for the analysis. Their basic characteristics are summarised in Table 5.1.

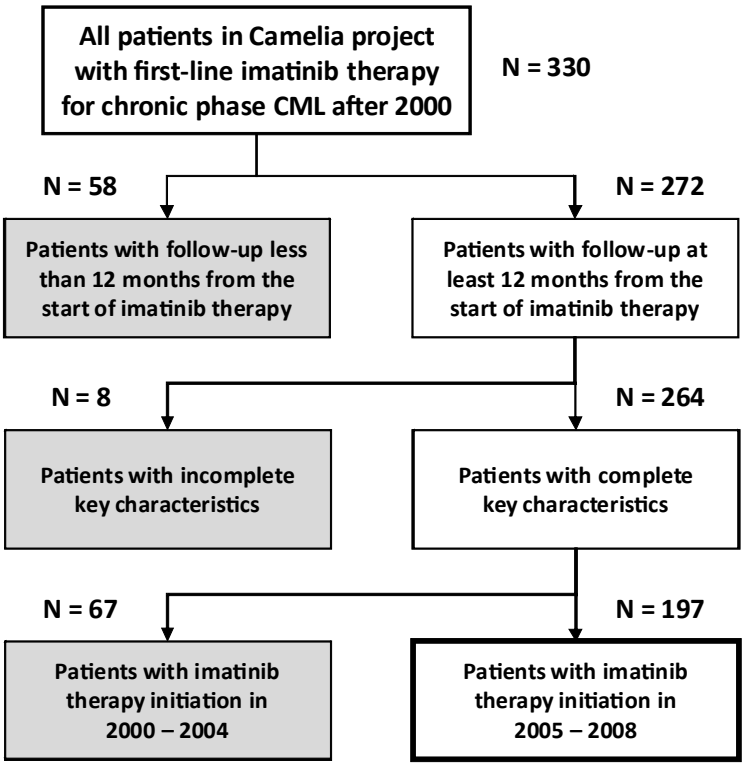

Figure 5.2: Definition of the final set of patients from Camelia project for Cox regression survival model.

5.3 Modelling the primary endpoint

A Cox proportional hazards model [20] was used to study the association between the time to the CCgR or the MMR to imatinib therapy and explanatory variables recorded in the Camelia project. Both fixed effects and random effects, the latter represented by the frailty term, were considered.

It should be noted that, with respect to achievement of the CCgR or the MMR as a primary endpoint of this study, there may occur competing events preventing the occurrence of the event of interest during the follow-up of the patient. Obviously, first of the competing events is death which causes the CCgR or the MMR not to occur any more. As a second competing event, a termination of the imatinib therapy can be regarded. Of course, the termination can have various reasons, of which undergoing stem cell transplantation, imatinib-related adverse events, and resistance to imatinib treatment are the best examples. However, the quantification of risk associated with the covariates with respect to these alter-

Table 5.1: Basic characteristics of patients with chronic myeloid leukemia considered in the analysis (N=197).

Characteristic	Level	N	%
Age in years	<50 years	96	48.7
	50+ years	101	51.3
Sex	Female	84	42.6
	Male	113	57.4
Sokal score	Low risk	78	39.6
	Intermediate risk	73	37.1
	High risk	46	23.4
Hasford score	Low risk	85	43.1
	Intermediate risk	86	43.7
	High risk	26	13.2
Achievement of CCgR or MMR	Yes	155	78.7
	No	42	21.3

native events was not of primary interest here, so the patients with competing events were treated as censored observations. A justification for this step was outlined in Section 2.3.7.

The multivariate exploratory analysis was performed using Statistica 9 [86] due to its visualization possibilities, whereas Cox regression model was fit with R software for statistical computing [78], using the `survival` package programmed by Terry Therneau.

5.3.1 Primary variable selection

Prior to Cox model construction, the continuous explanatory variables eligible for the modelling were analysed using a multivariate exploratory techniques, cluster analysis and principal component analysis [65], to identify highly correlated prognostic factors and, therefore, to avoid multicollinearity in the Cox model. Discrete explanatory variables were not considered for the clustering and their further inclusion to Cox regression model was judged individually. Moreover, some continuous variables were also dropped out from the list of variables feasible for the clustering procedure for high percentage of missing values due to objective reasons (e.g. variables recently added to parametric structure of the study, special examinations available only in part of the patient set). Continuous variables not distributed according to normal distribution, i.e. in particular the blood counts, were log transformed before the multivariate exploratory analysis was carried out.

There were four distinct clusters and two separate clinical variables identified with the multivariate techniques. The first cluster including hemoglobin is a cluster corresponding to red blood cells count and part of the blood differential. The second cluster represents patient's bone marrow status, i.e. the cluster corresponds to the bone marrow involvement with the disease. The third cluster containing age and prognostic scores is associated with overall health status of patients with respect to the CML burden. Finally, the fourth cluster constitutes of cytogenetic examinations mirroring the CML burden on the molecular level. The two clinical variables separated alone are MCV (*mean cell volume*, a measure of the average red blood cell volume), and leukocytes, i.e. the number of white blood cells, respectively.

Every time a Cox model was built up, only one member from each of the identified groups of prognostic factors was used as a covariate. In addition, following categorical variables were also considered for the modelling:

- *Patient's sex*. The inclusion of sex is due to the fact that sex was previously shown as a potential factor that may play role in a response to targeted therapy of non-small cell lung cancer. Male gender was taken as a risk category.

- *Imatinib dosage*. Dosage is a key point of almost every pharmacotherapy; four different dosing regimens were considered for imatinib in this study: 400 mg/day refers to a standard dose, >400 mg/day refers to an escalated dose, <400 mg/day refers to a reduced dose, and 0 mg/day refers to a temporary discontinued imatinib therapy, which is mainly due to mild side effects.

- *Clonal chromosomal abnormalities in the Ph+ cells*. The occurrence of clonal chromosomal abnormalities in the cells bearing the Philadelphia chromosome in time is known as a risk factor for development of a resistance to imatinib therapy and subsequent disease progression. Thus, the chromosomal abnormalities in Ph+ cells were also considered as a potential risk factor according to achievement of the CCgR or the MMR.

- *Clonal chromosomal abnormalities in the Ph- cells*. The clonal chromosomal abnormalities can occur also in other cell lines than those bearing the Philadelphia chromosome. Although these disorders are rather rare, they were also included in the model.

The variables mentioned so far were incorporated in the modelling process as fixed effects, however, two random effects were also considered to explain part of the variability present in the CML data. First, an univariate frailty model with each individual patient having his or her own frailty was fit to the data, whereas,

as the second model, a shared frailty model with each clinical centre having its own centre-specific frailty term was used.

5.3.2 Construction of the final model

Variables representing the clusters identified by multivariate methods were considered as time-fixed effects in the Cox model whereas imatinib dosage and clonal chromosomal abnormalities in the Ph+ and Ph- cells, respectively, were incorporated as time-varying effects. However, to fit a Cox model with time-varying covariates using the `survival` package requires the patient's data to be prepared in a requisite form, i.e. one row for each period of observation with a certain (non-missing) value of time-varying covariate. In this study, the length of one period was set to 1 month for monthly intervals are sufficient for CML monitoring. Since tied survival times were present in the observed data, Efron's approximation of partial log likelihood was used for estimation of the regression coefficients.

Table 5.2: Hazard ratios identified with Model 1 according to achievement of cytogenetic or molecular response to imatinib therapy in chronic CML patients treated with imatinib in first-line after 2004 (N=197).

Risk factor	Risk category / Basal category	Hazard ratio	95 % CI	p-value
Sex	Male / Female	1.63	1.15–2.32	0.006
Hemoglobin	Hb $<$ 110 g/l / Hb 110 g/l and more	1.53	1.00–2.33	0.051
MCV	MCV $>$ 100 fl / MCV 100 fl and less	2.19	1.24–3.86	0.007
Sokal score	Medium risk / Low risk	1.33	0.92–1.92	0.130
Sokal score	High risk / Low risk	2.29	1.36–3.87	0.002
Imatinib dosage	Lower dose / Standard or higher dose	7.53	2.64–21.51	$<$0.001
Clinical centre*	-	-	-	$<$0.001

* Included as a random effect

Since some of the factors were known *a priori* to play role in CML progression, the backward selection of variables was used to fit the Cox model. As sufficient statistical information need to be conveyed in the data for survival modelling [65], a rule of thumb, that there should be at least 10 times as many events as there are candidate covariates in the sample, was met with respect to the full model [49]. Furthermore, as a model fit can be highly influenced by specification errors in functional form of individual covariates [48], the correct functional form of continuous covariates was checked using spline approximation of individual covariates during the modelling process (the `survival` package allows for the testing of the nonlinear effect of a spline-smoothed covariate). On the other hand, the continuous covariates can be also categorised as there are threshold values with a clear clinical interpretation for some of the clinical characteristics, particularly

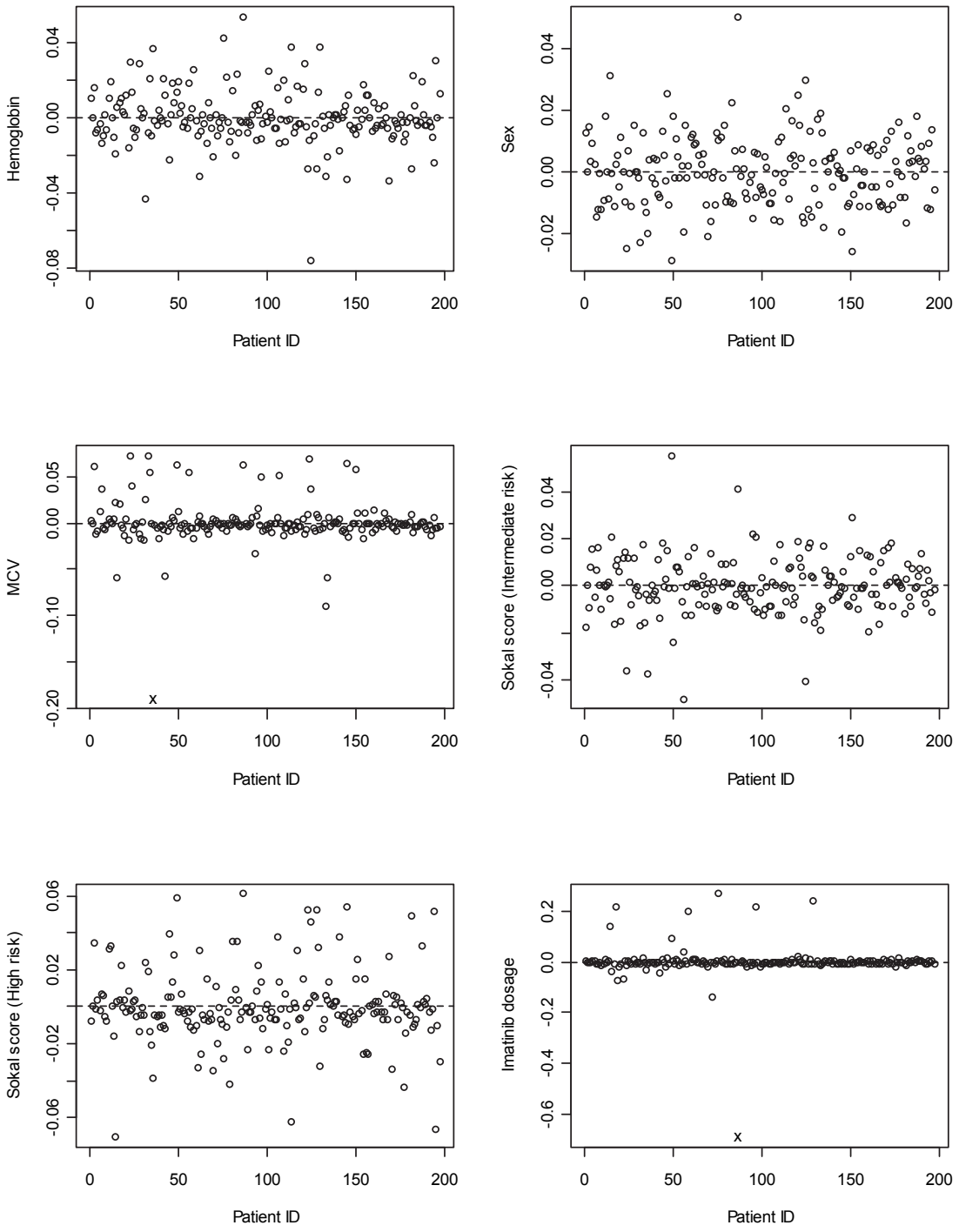

Figure 5.3: Plots of scaled score residuals according to covariates included in Model 1; the x-axis corresponds to individual patients (N=197).

blood count variables. For example, this is the case of hemoglobin and MCV, respectively. In the former case, value of 110 g/l was used to split the patients into those with and without anemia, whereas in the latter case, value of 100 fl was used to distinguish patients with macrocytic red blood cells (MCV>100 fl).

Reduction of the full model was performed gradually with the use of likelihood ratio test. As a result, model denoted as Model 1 was fit including following covariates: sex, categorised hemoglobin, categorised MCV, Sokal score, and imatinib dosage. The resulting hazard ratios associated with the set of covariates included in Model 1 are shown in Table 5.2 as well as their 95 % confidence intervals and corresponding p-values. No relevant interactions between the considered variables were identified. As a random effect, only the centre-specific frailty was identified as statistically significant, however, the choice of either gamma or inverse Gaussian frailty distribution had a negligible effect on the results.

Regression diagnostic was performed to find out whether Model 1 adequately describes the data. First, the set of variables was tested for the proportionality of hazards using a test based on weighted Schoenfeld residuals according to [40], and the proportionality of hazards was not rejected for any of the covariates as well as for the model as a whole. Second, scaled score residuals (also denoted as *dfbeta* residuals) were used to identify influential observations according to individual covariates, the resulting plots can be seen on Figure 5.3. Considering the magnitude of the regression coefficients, two highly influential observations can be seen on Figure 5.3, one for MCV, and one for imatinib dosage (both marked with "x"s). Omitting these two observations from data yielded a model with the regression coefficient for MCV being slightly shrink towards zero, the regression coefficient for imatinib dosage being strongly shrink towards zero, while the rest of the coefficients being approximately the same (results not shown here). Moreover, examination of the scaled score residuals once again revealed three highly influential observations, again associated with MCV and imatinib dosage. Fitting a new model without the three influential observations resulted in a model showing no statistical significance for MCV and imatinib dosage any more. Based on this fact, model denoted as Model 2 was fit including only sex, categorised hemoglobin, Sokal score, and clinical centre as a random effect, in the model formula.

Summary of Model 2 is shown in Table 5.3 showing the estimated hazard ratios and their 95 % confidence intervals. Regression diagnostic was also performed for Model 2. First, there is no evidence of non-proportional hazards both globally and individually for all four included variables. Second, none of the observations is heavily influential individually with respect to the magnitudes of the regression coefficients, see Figure 5.4. Third, a fairly good fit of Model 2 is indicated by a plot of deviance residuals, see Figure 5.5, that are scattered

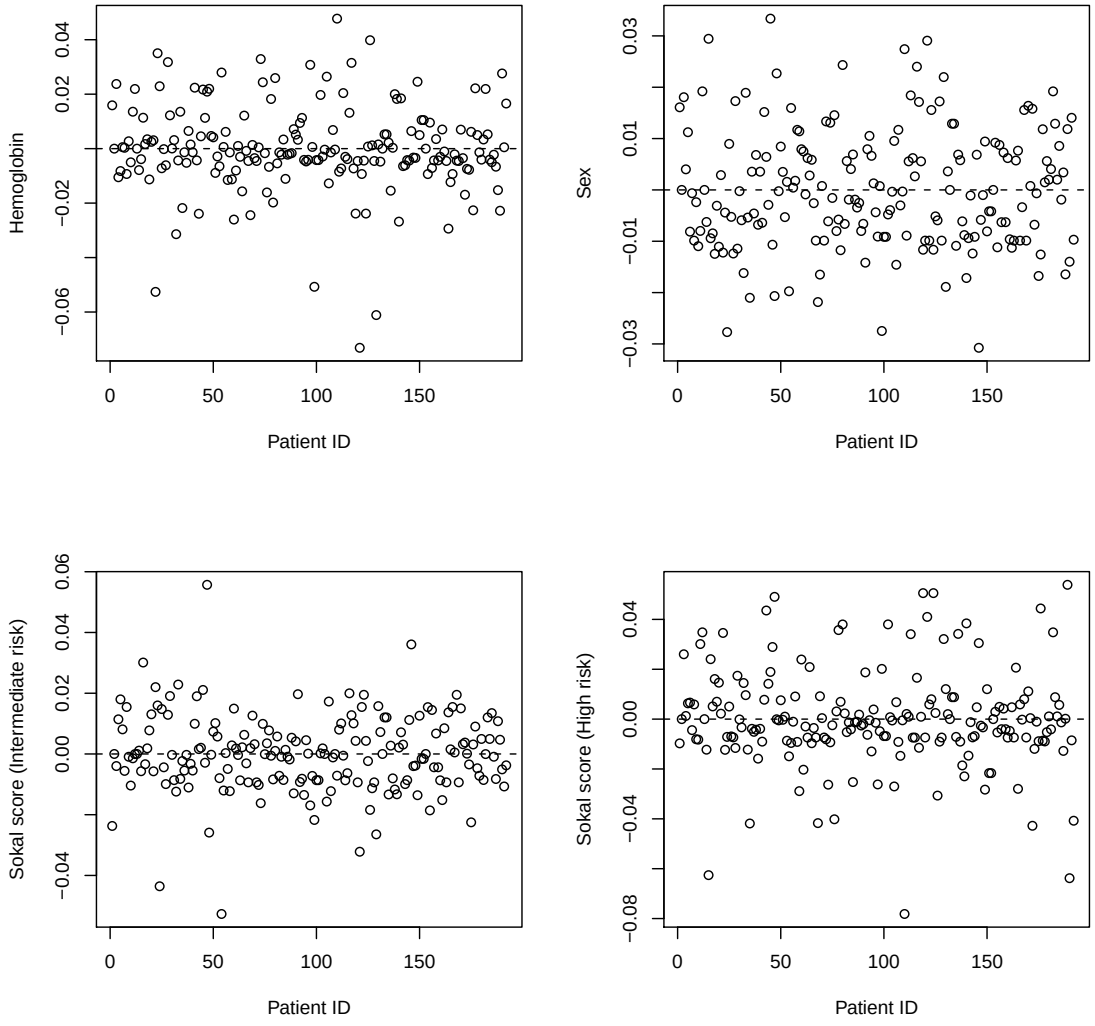

Figure 5.4: Plots of scaled score residuals according to covariates included in Model 2; the x-axis corresponds to individual patients (N=192).

Table 5.3: Hazard ratios identified with Model 2 according to achievement of cytogenetic or molecular response to imatinib therapy in chronic CML patients treated with imatinib in first-line after 2004 (N=192).

Risk factor	Risk category / Basal category	Hazard ratio	95 % CI	p-value
Sex	Male / Female	1.88	1.33–2.66	<0.001
Hemoglobin	Hb < 110 g/l / Hb 110 g/l and more	1.89	1.23–2.87	0.004
Sokal score	Medium risk / Low risk	1.34	0.93–1.93	0.120
Sokal score	High risk / Low risk	2.43	1.45–4.08	<0.001
Clinical centre*	-	-	-	<0.001

* Included as a random effect

about zero with no observations being far from a hypothetical horizontal line at zero [93]. This finding was also supported with a test of overall goodness-of-fit proposed by Parzen and Lipsitz for the Cox model [73]. The value of their score statistic based on eight risk categories was 3.66 with p = 0.183 (df = 7). Thus, the hypothesis of model fit cannot be rejected, at a significance level of 0.05.

5.4 Discussion

In this chapter, a Cox regression model was built up for chronic phase CML patients treated with imatinib in first-line, that aimed to identify the potential risk factors associated with prolonged time to achievement of the CCgR or the MMR to imatinib therapy. The model was based on patients coming from the Camelia project which is an observational study of 10 Czech and Slovak clinical centres. Due to consistency of patient data, total of N=197 prospective patients administered to imatinib therapy in 2005–2008 were considered for the modelling.

The issue of multicollinearity was addressed with reduction of the set of continuous explanatory variables using the cluster analysis and principal component analysis. The so-called univariate screening method was not used as this method does not account for the joint nature of the problem. The multivariate methods identified several clinically relevant clusters, each of which was further represented in the Cox model with one variable. In addition, four fixed effects corresponding to sex, imatinib dosage, Ph+ and Ph- clonal chromosome abnormalities, respectively, and one random effect corresponding either to individual heterogeneity or centre-specific heterogeneity were considered in the full model. The backward selection of covariates was used to fit the Cox model with likelihood ratio test serving as a tool for eliminating of insignificant variables. Moreover, regression diagnostic techniques were applied to check for the model adequacy.

Model with sex, hemoglobin, Sokal score as fixed effects, and clinical centre as

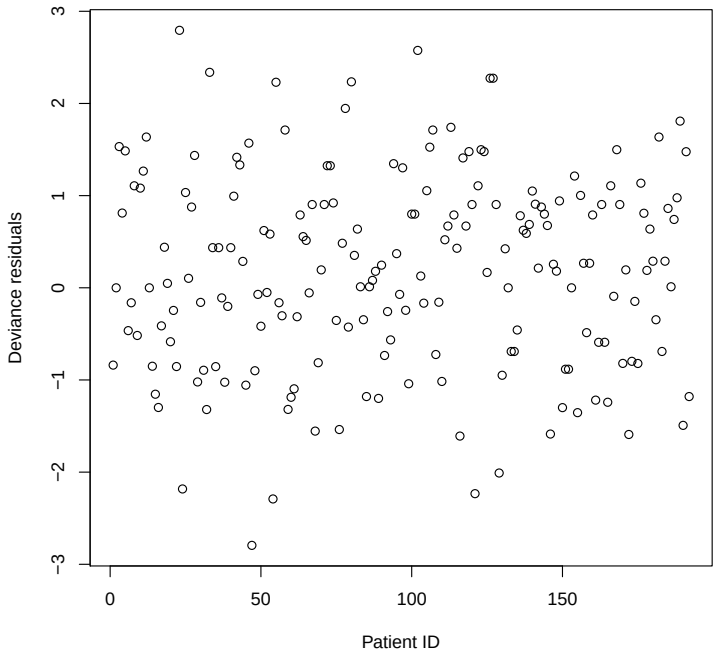

Figure 5.5: Deviance residuals for Model 2 plotted against individual patients (N=192).

random effect was chosen as the final model for the achievement of the CCgR or the MMR to imatinib therapy. The statistical significance of shared frailty term representing the individual clinical centres is not surprising for shared frailty models have been commonly used for survival assessment in multi-centre studies, especially in those where some centres contain only several patients [99]. The prognostic potential of the Sokal score, which seems to be still valid, is also not surprising, even if Sokal score [83] was derived in time, when imatinib was not yet in clinical use. It should be also noted that substitution of Sokal score with Hasford score leads only to minor changes is parameter estimates. However, the insignificant difference in hazard profiles between low risk and intermediate risk indicates that there is a space for a new prognostic score to be developed for the *era of imatinib therapy*, or more precisely for the *era of targeted therapy*. Anemia has been known for long as one of the symptoms of CML as well as one of the factors associated with worse prognosis [57], therefore the association between anemia and an elevated risk of delayed or no achievement of the CCgR or the MMR to imatinib therapy is clinically plausible. On the other hand, such a difference in the achievement of the CCgR and the MMR to imatinib therapy between males and females has not been published in CML so far, which raises a question whether it

could be explained with some kind of administrative issues as in case of period of diagnosis described above. Another possible explanation of this association may be different compliance of both sexes to imatinib therapy. As the introduction of imatinib has revolutionized the treatment of CML, and many patients may have almost negligible CML burden, patient compliance to imatinib therapy may play a role in the achievement of the CCgR or the MMR.

The lack of statistical significance of clonal chromosome abnormalities, which are known as risk factors for CML progression, can be explained with insufficient statistical information for there is only a small number of patients who developed Ph+ or Ph- clonal chromosome alterations. The same can be also true for the imatinib dosage which was eliminated from the final model after omitting two influential observations, both with long response-free period and reduced imatinib dose at the same time. It is obvious from the clinical perspective that lower than standard dose of imatinib should be associated with limited response, however, more observations with reduced imatinib dose would be needed to support this hypothesis statistically.

It should be noted that even if the covariates in Cox regression are statistically significant and regression diagnostic tools show no problem with the fit of the observed data, the predictive power of the model can be limited at the individual level, i.e. with respect to individual patient. Obviously, this is even more likely to be true if the number of observed failures is not sufficiently large, and the model is not stable enough according to covariates included in the model. Therefore, considering Model 2 presented in Table 5.3, its potential in identification of patients, who are more likely to have problems with proper treatment response to imatinib therapy, should be emphasized rather than its potential in prediction of precise time to achievement of the CCgR or the MMR.

6 Conclusion

In this dissertation, methods for the assessment and modelling in the field of cancer epidemiology and their applications on real data sets were focused. A new model for the estimation of prevalence of patients requiring active anti-tumour therapy was presented. First advantage of this model is that it utilizes only population-based cancer registry data. Second advantage is that, unlike the methods published so far, the new model allows the prevalence to be estimated with respect to the extent of cancer which is for many types of cancer more important than age at diagnosis. The applicability of the model was shown on colorectal cancer data from the Czech National Cancer Registry to model the number of potentially treated patients with colorectal carcinoma in the Czech Republic in 2011. Moreover, the survival benchmarks for the Czech cancer patients in the form of five-year relative survival rates were presented. These estimates can be readily used for health care assessment on the population level. Also an overview of the five-year relative survival rates achieved since 1990 was provided as well as a comparison between the Czech and European five-year relative survival rates in selected cancer diagnoses. Finally, the applicability of survival regression techniques on the data from an international registry for monitoring of incidence, treatment and treatment response in patients with chronic myeloid leukemia in the Czech Republic and Slovakia was shown. A Cox regression model was built up for patients in chronic phase of chronic myeloid leukemia treated with imatinib in first-line, that aimed to identify the potential risk factors associated with prolonged time to achievement of the complete cytogenetic or the major molecular response to modern targeted therapy.

References

[1] Aalen OO (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6: 701726.

[2] Baccarani M, Saglio G, Goldman J, et al. (2006). Evolving concepts in the management of chronic myeloid leukemia: recommendations from an expert panel on behalf of the European LeukemiaNet. *Blood*, 108(6): 1809-1820.

[3] Baccarani M, Cortes J, Pane F, et al. (2009). Chronic myeloid leukemia: an update of concepts and management recommendations of European LeukemiaNet. *Journal of Clinical Oncology*, 27(35): 6041-6051.

[4] Barker PN & Henderson R (2005). Small sample bias in the gamma frailty model for univariate survival, *Lifetime Data Analysis*, 11: 265284.

[5] Bashir S & Estve J (2001). Projecting cancer incidence and mortality using Bayesian age-period-cohort models, *Journal of Epidemiology and Biostatistics*, 6: 287-296.

[6] Berrino F, DeAngelis R, Sant M et al. (2007). Survival for eight major cancers and all cancers combined for European adults diagnosed in 199599: results of the EUROCARE-4 study. *Lancet Oncology*, published online in August 21, 2007: http://oncology.thelancet.com.

[7] Berry DA, Cronin KA, Plevritis SK et al. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England Journal of Medicine*, 353: 17841792.

[8] Bray F & Møller B (2006). Predicting the future burden of cancer, *Nature Reviews Cancer*, 6: 63-74.

[9] Brenner H & Hakulinen T (2002). Up-to-date long-term survival estimates of patients with cancer by period analysis. *Journal of Clinical Oncology*, 20: 826-832.

[10] Brenner H & Hakulinen T (2003). On crude and age-adjusted relative survival rates. *Journal of Clinical Epidemiology*, 56(12): 1185-1191.

[11] Brenner H, Gefeller O & Hakulinen T (2004). Period analysis for up-to-date cancer survival data: Theory, empirical evaluation, computational realization and applications. *European Journal of Cancer*, 40: 326-335.

[12] Brenner H, Gondos A & Arndt V (2007). Recent major progress in long-term cancer patient survival disclosed by modeled period analysis. *Journal of Clin-*

ical Oncology. 25(22): 32743280.

[13] Breslow NE (1972). Discussion of paper of D. R. Cox. *Journal of the Royal Statistical Society, Series B*, 34: 216-217.

[14] Breslow NE & Day NE (1980). *Statistical Methods in Cancer Research. Volume I - The Analysis of Case-Control Studies*. Lyon, International Agency for Research on Cancer (IARC Scientific Publications No. 32).

[15] Capocaccia R & De Angelis R (1997). Estimating the completeness of prevalence based on cancer registry data, *Statistics in Medicine*, 16: 425-440.

[16] 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.

[17] Clegg LX, Gail MH, Feuer EJ (2002). Estimating the variance of disease-prevalence estimates from population-based registries, *Biometrics*, 58: 684-688.

[18] Coleman M, Quaresma M, Berrino F, Lutz J, DeAngelis R, Capocaccia R, Baili P, Rachet B, Gatta G, Hakulinen T et al. (2008). Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncology*, 9(8): 730756.

[19] Corazziari I, Quinn M, Capocaccia R (2004). Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer*, 40: 23072316.

[20] Cox D (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34: 187-220.

[21] Czech Oncology in Numbers.

[22] De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A (1999). Mixture models for cancer survival analysis: Application to population-based data with covariates, *Statistics in Medicine*, 18: 441-454.

[23] Dickman PW & Hakulinen T (2003). Population-based cancer survival analysis. Available from URL http://www.pauldickman.com/teaching/-tampere2004/book_draft.pdf.

[24] Dickman PW, Sloggett A, Hills M & Hakulinen T (2004). Regression models for relative survival. *Statistics in Medicine*, 23(1): 5164.

[25] Dickman PW & Adami HO (2006). Interpreting trends in cancer patient survival. *Journal of Internal Medicine*, 260(2): 103117.

[26] dos Santos Silva I (1999). *Cancer Epidemiology: Principles and Methods*, Lyon, International Agency for Research on Cancer (IARC Nonserial Publication).

[27] Dušek L, Pavlík T, Abrahámová J, Koptíková J, Mužík J, Gelnarová E (2008). Referenční data pro hodnocení léčebné péče u zhoubných nádorů varlat. In *Nádory varlat*, Praha, Grada Publishing.

[28] Dušek L, Pavlík T, Májek O, Koptíková J, Gelnarová E, Mužík J, Vyzula R,

Fínek J (2009). Information System for Predictive Evaluation of Cancer Epidemiology and the Number of Cancer Patients in the Czech Republic. In *Czech Cancer Care in Numbers 2008-2009*, Praha, Grada Publishing.

[29] Dyba T & Hakulinen T (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques, *Statistics in Medicine*, 19(13): 1741-1752.

[30] Dyba T & Hakulinen T (2008). Do cancer predictions work? *European Journal of Cancer*, 44(3): 448-453.

[31] Ederer F & Heise H (1959). Instructions to IBM 650 programmers in processing survival computations. *Methodological note No. 10*, End Results Evaluation Section, National Cancer Institute, Bethesda MD.

[32] Elbers C & Ridder G (1982). True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model. *The Review of Economic Studies*, 49(3): 403-409.

[33] Feldman AR, Kessler L, Myers MH, Naughton MD (1986). The prevalence of cancer: estimates based on Connecticut Cancer registry, *The New England Journal of Medicine*; 315: 1394-1397.

[34] Ferlay J, Autier P, Boniol M, Heanue M, Colombet M, Boyle P (2007). Estimates of the cancer incidence and mortality in Europe in 2006. *Annals of Oncology*, 18(3): 581-592.

[35] 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.

[36] 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.

[37] Goldman JM, Melo JV (2003). Chronic myeloid leukemia–advances in biology and new approaches to treatment. *The New England Journal of Medicine*, 349(15): 1451-1464.

[38] Gondos A, Arndt V, Holleczek B, Stegmaier C, Ziegler H & Brenner H (2007). Cancer survival in Germany and the United States at the beginning of the 21st century: an up-to-date comparison by period analysis. *International Journal of Cancer*. 121(2): 395400.

[39] Gondos A, Bray F, Brewster DH, Coebergh JW, et al. (2008). Recent trends in cancer survival across Europe between 2000 and 2004: a model-based period analysis from 12 cancer registries. *European Journal of Cancer*. 44(10): 14631475.

[40] Grambsch PM & Therneau TM (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, 81: 515-527.

[41] Grande E, Inghelmann R, Francisci S, Verdecchia A, Micheli A, Baili P,

Capocaccia R, De Angelis R (2007). Regional estimates of colorectal cancer burden in Italy. *Tumori*, 93(4): 352-359.

[42] Gras C, Daurs JP, Tretarre B (2004). Age and stage specific prevalence estimate of cancer from population based Cancer Registry using inhomogeneous Poisson process, *Statistical Methods in Medical Research*, 13: 273-289.

[43] Greenwood M (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*. London: Her Majesty's Stationery Office; 33: 126.

[44] Gutierrez RG (2002). Parametric frailty and shared frailty survival models. *The Stata Journal*, 2(1): 22-44.

[45] Hakulinen T (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38: 933942.

[46] Hakulinen T, Teppo L, Saxen E (1986). Do the predictions for cancer incidence come true? Experience from Finland, *Cancer*, 57(12): 2454-2458.

[47] Hakulinen T & Dyba T (1994). Precision of incidence predictions based on Poisson distributed observations, *Statistics in Medicine*, 13: 1513-1523.

[48] Harrell FE Jr (2001). *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*, New York, Springer.

[49] Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA (1984). Regression modelling strategies for improved prognostic prediction, *Statistics in Medicine*, 3(2): 143-152.

[50] Hasford J, Pfirrmann M, Hehlmann R, et al. (1998). A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. Writing Committee for the Collaborative CML Prognostic Factors Project Group, *Journal of the National Cancer Institute*, 90(11): 850-858.

[51] Heinavaara S & Hakulinen T (2006). Predicting the lung cancer burden: Accounting for selection of the patients with respect to general population mortality, *Statistics in Medicine*, 25:2967-2980.

[52] Holford TR (1983). The estimation of age, period and cohort effects for vital rates, *Biometrics*, 39(2): 311-324.

[53] Hosmer DW, Lemeshow S, May S (2008). *Applied survival analysis: regression modeling of time to event data*, New York, John Wiley & Sons.

[54] Ito Y, Ohno Y, Rachet B, Coleman MP, Tsukuma H, Oshima A (2007). Cancer survival trends in Osaka, Japan: the influence of age and stage at diagnosis. *Japanese Journal of Clinical Oncology*. 37(6): 452458.

[55] Janssen-Heijnen ML, Houterman S, Lemmens VE, Brenner H, Steyerberg EW, Coebergh JW (2007). Prognosis for long-term survivors of cancer. *Annals of Oncology*. 18(8): 14081413.

[56] Kalbfleisch JD & Prentice RL (2002). *The Statistical Analysis of Failure Time Data*, New York, John Wiley & Sons.

[57] Kantarjian HM, Smith TL, McCredie KB, et al. (1985). Chronic myelogenous

leukemia: a multivariate analysis of the associations of patient characteristics and therapy with survival, *Blood*, 66(6): 1326-1335.

[58] Kaplan EL & Meier P (1958). Nonparametric estimation from incomplete observations, *Journal of American Statistical Association*, 53: 457-481.

[59] Klein JP (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm, *Biometrics*, 48(3): 795-806.

[60] Klein JP & Moeschberger ML (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York, Springer.

[61] Kruijshaar ME, Barendregt JJ, van de Poll-Franse LV (2003). Estimating the prevalence of breast cancer using a disease model: data problems and trends, *Population Health Metrics*, 1:5.

[62] Lambert PC, Thompson JR, Weston CL, Dickman PW (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8: 576-594.

[63] Lunn M & McNeil D (1995). Applying Cox regression to competing risks, *Biometrics*, 51(2): 524-532.

[64] Mandel JS, Church TR, Bond JH, Ederer F, Geisser MS, Mongin SJ, Snover DC, Schuman LM (2000). The effect of fecal occult-blood screening on the incidence of colorectal cancer. *The New England Journal of Medicine*, 343(22): 1603-1607.

[65] Marubini E & Valsecchi MG (2004). *Analysing Survival Data from Clinical Trials and Observational Studies*. New York, John Wiley & Sons.

[66] Møller B, Fekjaer H, Hakulinen T, Sigvaldason H, Storm HH, Talback M, Haldorsen T (2003). Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. *Statistics in Medicine*, 22(17): 2751-2766.

[67] Mužík J, Koptíková J, Dušek L, Žaloudík J, Vyzula R, Abrahámová J (2007). Historical data of Czech National Cancer Registry: information value and risk of bias. *Klinická onkologie*, 20(Suppl. 1): 63-76.

[68] Nelson W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1: 2752.

[69] Neuss MN, Desch CE, McNiff KK, Eisenberg PD et al. (2005). A Process for Measuring the Quality of Cancer Care: The Quality Oncology Practice Initiative. *Journal of Clinical Oncology*, 23(25): 62336239.

[70] Nielsen GG, Gill RD, Andersen PK, Sørensen TIA (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics*, 19(1):25-43.

[71] Parkin DM, Whelan SL, Ferlay J, Storm H (2005). *Cancer Incidence in Five Continents*, Vol. I to VIII IARC CancerBase No. 7, Lyon.

[72] Parner E (1997). Inference in semiparametric frailty models. Technical report,

Ph.D. dissertation, University of Aarhus, Denmark.

[73] Parzen M, Lipsitz SR (1999). A global goodness-of-fit statistic for Cox regression models. *Biometrics*, 55(2): 580-584.

[74] Pavlík T, Dušek L, Májek O, Žaloudík J (2009). Five-Year Survival Rates of Cancer Patients in the Czech Republic. In *Czech Cancer Care in Numbers 2008-2009*, Praha, Grada Publishing.

[75] Pavlík T, Dušek L, Májek O, Koptíková J, Vyzula R, Fínek J (2009). Modelling number of cancer patients potentially treated with targeted pharmacotherapy in the czech republic. In *International Society for Pharmacoeconomics and Outcomes Research*, International Society for Pharmacoeconomics and Outcomes Research, Value in Health, 12(7): A287.

[76] Phillips N, Coldman A, McBride ML (2002). Estimating cancer prevalence using mixture models for cancer survival. *Statistics in Medicine*, 21(9): 1257-1270.

[77] Quintás-Cardama A, Kantarjian H, Jones D, Shan J, Borthakur G, Thomas D, Kornblau S, O'Brien S, Cortes J (2009). Delayed achievement of cytogenetic and molecular response is associated with increased risk of progression among patients with chronic myeloid leukemia in early chronic phase receiving high-dose or standard-dose imatinib therapy. *Blood*, 113(25): 6315-6321.

[78] R Development Core Team (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[79] Rachet B, Woods LM, Mitry E, Riga M et al. (2008). Cancer survival in England and Wales at the end of the 20th century. *British Journal of Cancer*. 99(Suppl 1): S210.

[80] Rashid I, Marcheselli L & Federico M (2008). Estimating survival in newly diagnosed cancer patients: Use of computer simulations to evaluate performances of different approaches in a wide range of scenarios, *Statistics in Medicine*, 27: 21452158.

[81] Schoenfeld D (1982). Partial residuals for the proportional hazards model. *Biometrika*, 69: 239-241.

[82] Simonetti A, Gigli A, Capocaccia R, Mariotto A (2008). Estimating complete prevalence of cancers diagnosed in childhood, *Statistics in Medicine*, 27(7): 990-1007.

[83] Sokal JE, Cox EB, Baccarani M, et al. (1984). Prognostic discrimination in "good-risk" chronic granulocytic leukemia, *Blood*, 63(4): 789-799.

[84] Sposto R (2002). Cure model analysis in cancer: An application to data from the Children's Cancer Group. *Statistics in Medicine*, 21: 293-312.

[85] StataCorp, Inc. (2008). Stata statistical software: release 10. College Station,

TX: StataCorp LP.
[86] StatSoft, Inc. (2009). STATISTICA (data analysis software system), version 9.0. www.statsoft.com.
[87] Dusek L, Muzik J, Kubasek M, Koptikova J, Zaloudik J, Vyzula R (2005). Epidemiology of malignant tumours in the Czech Republic [online]. Masaryk University, [cit. 2009-12-15]. On-line available: http://www.svod.cz. Version 7.0 [2007], ISSN 1802-8861.
[88] Tabata N, Ohno Y, Matsui R, Sugiyama H, Ito Y, Tsukuma H, Oshima A (2008). Partial cancer prevalence in Japan up to 2020: estimates based on incidence and survival data from population-based cancer registries, *Japanese Journal of Clinical Oncology*, 38(2): 146-57.
[89] Talbäck M, Rosén M, Stenbeck M, Dickman PW (2004). Cancer patient survival in Sweeden at the beginning of the third millenium predictions using period analysis. *Cancer Causes and Control*, 15: 967976.
[90] Therneau TM & Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*, New York, Springer.
[91] Therneau TM, Grambsch PM, Pankratz VS (2000). Penalized survival models and frailty, Technical Report Series Number 66, Department of Health Sciences Research, Mayo Clinic, Rochester, USA.
[92] Therneau TM, Grambsch PM, Pankratz VS (2003). Penalized survival models and frailty, *Journal of Computational and Graphical Statistics*, 12(1): 156-175.
[93] Thompson LA, Chhikara RS, Conkin J (2003). Cox proportional hazards models for modeling the time to onset of decompression sickness in hypobaric environments, *NASA Technical Publication* 2003-210791, Houston: Johnson Space Center.
[94] Verdecchia A, De Angelis G, Capocaccia R (2002). Estimation and projection of cancer prevalence from cancer registry data, *Statistics in Medicine*, 21: 3511-3526.
[95] Verdecchia A, Francisci S, Brenner H et al. (2007). Recent cancer survival in Europe: a 2000–2002 period analysis of EUROCARE-4 data. *Lancet Oncology*, published online in August 21, 2007: http://oncology.thelancet.com.
[96] Visser O & van Leeuwen FE (2005). Stage-specific survival of epithelial cancers in North-Holland/Flevoland, The Netherlands, *European Journal of Cancer*, 41: 2321-2330.
[97] Wang JL (2005). Smoothing hazard rate. *Encyclopedia of Biostatistics*, 2nd Edition, Vol 7, 4986-4997.
[98] Welch HG, Schwartz LM, Woloshin S (2000). Are incresing 5-year survival rates evidence of success against cancer? *Journal of the American Medical Association*, 283: 29752978.
[99] Wintrebert CMA, Putter H, Zwinderman AH, van Houwelingen JC (2004).

Centre-effect on survival after bone marrow transplantation: application of time-dependent frailty models, *Biometrical Journal*, 46(5): 512-525.

List of figures

List of tables