Univerza
v Ljubljani
Medicinska
fakulteta

Klemen Pavlič

Ocenjevanje in primerjava krivulj čistega preživetja

Estimation and comparison of net survival curves

Doktorsko delo

Imenovanje mentorja na seji senata dne 14.6.2016
Komisija za spremljanje doktorskega študenta imenovana na seji senata dne 7.3.2016

Datum zagovora: 10.4.2018

Mentor: izr. prof. dr. Maja Pohar Perme

Predsednik komisije: prof. dr. Janez Stare
Član: prof. dr. Mihael Perman
Član: prof. dr. Per Kragh Andersen

Zahvala

Zahvaljujem se mentorici Maji Pohar Perme za vodenje pri raziskovalnem delu, za spodbudne besede v trenutkih, ko mi ni šlo in za predstavitev zanimivih problemov ter neštete diskusije o njih. Še posebej pa sem hvaležen za podporo pri tem, da se lahko ukvarjam s problemi, ki se mi zdijo zanimivi. Hvala predstojniku prof. Janezu Staretu za spodbujanje raziskovalnega vzdušja na inštitutu in za vse ideje s kolesa. Hvala tudi vsem ostalim sodelavcem z inštituta, vsak je prispeval kak kamenček k temu mozaiku.

Special thanks also to prof. Per Kragh Andersen and prof. Torben Martinussen for giving me the opportunity to work with them during my short visit to the Section of Biostatistics in Copenhagen. It was a great experience and it opened my eyes.

Abstract

Objectives

Relative survival analyses is a subfield of survival analysis that deals with competing risks data where the cause of death information is unavailable or unreliable. It is typically used to investigate survival of cancer patients and several measures are estimated to achieve this. For all of them the population mortality tables are used as an external source of information that substitutes the missing cause of death information.

Net survival measure is particularly popular, it does not depend on the hazard of dying from other causes by definition. However, the definition itself relies on assumptions that cannot be tested from the data and this makes the use of this measure questionable.

The goal of this work is threefold. First, to propose a new measure that does not rely on these questionable assumptions and to investigate its association with net survival. Secondly, to provide a new approach to estimation of this measure and explore its properties. Thirdly, to investigate the methods for comparison of net survival between different groups and to compare them.

Hypotheses

The new measure presented in this work relies on fewer assumptions and has an interpretation that does not depend on them. It provides an alternative to the net survival measure.

The estimation approach proposed in this work is simple to understand and easy to implement in different statistical packages. It allows extensions to discretely measured time and has desirable properties.

The log-rank type test and the test of a coefficient from the additive regression model are comparable. They respond to the same alternatives.

Methods

The properties of the new measure, the new estimator and the comparison of methods for comparison of net survival curves between groups are analysed theoretically, by simulations and illustrated on real data. We try to draw parallels to classical survival whenever possible. We use pseudo observations to construct the estimator of the new measure. We use counting processes and martingales for the derivation of variance of the new estimation approach. In the last part, we build on the known association between the Cox model and log-rank test.

Results

Pseudo observations are used as estimators of individual quantities in the construction of the new estimator which is their first usage outside regression modelling.

The proposed estimator of the new measure gives practically identical results as the PP estimator of net survival, the only difference being the interpretation where the new measure requires fewer assumptions.

The estimators of the variance of the new estimator work well. The precise formula gives coverages close to the nominal level whereas the approximate formula slightly underestimates the variance.

The new estimation approach works well with discretely measured time. It has smaller bias than the PP estimator and coverages closer to the nominal level. As such, it outperforms the PP estimator when used with wider intervals.

The log-rank type test and the test of coefficient from the additive regression model are not identical but they behave in the same way. They both respond to the same alternatives and perform equally poorly against crossing hazard alternatives.

We implemented the log-rank type test in the `R` package `relsurv` [24].

Izvleček

Namen dela

Analiza relativnega preživetja je podpodročje analize preživetja, ki se ukvarja s sotveganji, ko razlog smrti ni znan ali ni zanesljiv. Najpogosteje se uporablja pri analizi preživetja bolnikov z rakom. Obstaja več mer, ki jih lahko poročamo. Pri ocenjevanju teh mer manjkajočo informacijo o razlogu smrti nadomestimo z informacijo o tveganju za smrt iz drugih vzrokov, ki jo dobimo iz populacijskih tabel umrljivosti.

Znotraj področja analize relativnega preživetja se pogosto poroča mera imenovana čisto preživetje. Ta po definiciji ni odvisna od tveganja za smrt iz drugih vzrokov, vendar pa njena definicija temelji na predpostavkah, ki jih ne moremo testirati na podatkih. Uporaba te mere je zato vprašljiva.

To delo ima tri glavne cilje. Prvi je definirati novo mero, ki ne bo temeljila na nepreverljivih predpostavkah, ki so potrebne za definicijo čistega preživetja, in raziskati povezavo med novo mero in čistim preživetjem. Drugi cilj je predlagati cenilki za novo mero in njeno varianco ter raziskati njune lastnosti. Zadnji cilj je raziskati in primerjati metode za primerjavo krivulj čistega preživetja.

Hipoteze

Nova mera, ki je definirana v tem delu, temelji na manj predpostavkah kot čisto preživetje. Kot taka nudi alternativo čistemu preživetju z nesporno interpretacijo.

Predlagana cenilka je enostavno razumljiva, saj izhaja neposredno iz definicije mere, prav tako pa omogoča enostavno implementacijo mere v različne programske pakete. Nudi tudi razširitev na podatke z diskretno merjenim časom in ima željene lastnosti.

Test oblike log-rank je primerljiv s testom koeficienta iz aditivnega regresijskega modela. Odzivata se na iste alternativne hipoteze.

Metode

Lastnosti nove mere, njene cenilke in metode za primerjavo krivulj čistega preživetja bomo raziskali teoretično in s simulacijami ter jih ilustrirali na pravih podatkih. Vlekli bomo vzporednice z znanimi rezultati iz analize preživetja. Za konstrukcijo nove cenilke bomo uporabili psevdo vrednosti, za konstrukcije cenilke njene variance pa si bomo pomagali z martingali in procesi štetja. Pri zadnjem delu bomo izhajali iz znane zveze med testom log-rank in testom koeficienta iz Coxovega modela.

Rezultati

Predlagana cenilka nove mere daje praktično identične rezultate kot cenilka PP za čisto preživetje. To kaže, da imata nova mera in čisto preživetje enake ocenjene vrednosti, razlika je le v interpretaciji. Pri tem ta pri novi meri temelji na manj predpostavkah.

Predlagali smo dve cenilki za ocenjevanje variance nove cenilke. Prva je natančna in daje tudi pokritja, ki se ujemajo z nominalno vrednostjo, druga pa je aproksimativna in rahlo podceni varianco, a so razlike v primerih, ki jih lahko pričakujemo v praksi, minimalne.

Novo cenilko smo razširili tudi na podatke z diskretno merjenim časom, kjer ima manjšo pristranskost kot cenilka PP, prav tako pa daje tudi pokritja, ki so bližje nominalni vrednosti.

Test oblike log-rank in test koeficienta iz aditivnega regresijskega modela nista identična, vendar se obnašata podobno. Odzivata se na iste alternativne hipoteze in sta podobno neobčutljiva za nesorazmerna tveganja.

Test oblike log-rank smo dodali v `R` paket `relsurv` [24].

Contents

1 Introduction **1**

1.1 The field of relative survival 1

1.1.1 Population mortality tables 1

1.1.2 Different measures 2

1.2 Goals of this work 4

1.3 Notation 5

1.4 Simulations and real data 6

1.4.1 Simulations 6

1.4.2 Data 7

2 The measure of interest **9**

2.1 Additive model 9

2.2 Latent times and definition of net survival 10

2.3 Marginal relative survival 11

2.4 Underlying assumptions and issues 12

2.4.1 Availability of data on excess term 12

2.4.2 Assumptions affecting comparability between populations . . . 14

2.4.3 Assumptions affecting interpretation . . . 15

3 Estimation of marginal relative survival 19

3.1 Pseudo observations . . . 20

3.1.1 Properties of pseudo observations for survival probability 22

3.2 Variance derivation and estimation . . . 25

3.2.1 Variance derivation . . . 25

3.2.2 Variance estimation . . . 27

3.3 Comparison of the PP estimator and the pseudo based approach 29

3.3.1 The PP estimator . . . 30

3.3.2 Comparison in theory . . . 30

3.3.3 Simulation study . . . 32

3.3.4 Example . . . 36

3.4 Discretely measured time . . . 38

3.4.1 Life table estimator . . . 39

3.4.2 Estimation of marginal relative survival and estimation of the variance of its estimator . . . 39

3.4.3 Simulation study . . . 40

4 Comparison of net survival curves 45

4.1 Methods for comparison . . . 45

4.1.1 Properties and interpretation of the log-rank type test 48

4.1.2 Simulations . . . 50

4.1.3 Simulation results 53

4.1.4 Example 61

5 Conclusions **65**

A Povzetek dela **69**

A.1 Uvod 69

A.2 Čisto in robno relativno preživetje 70

A.3 Ocenjevanje robnega relativnega preživetja 72

A.4 Primerjava krivulj čistega preživetja 74

A.5 Zaključek 75

Bibliography **77**

Chapter 1

Introduction

1.1 The field of relative survival

'Would it save you a lot of time if I just gave up and went mad now?'

Douglas Adams

The field of relative survival analysis is a subfield of survival analysis that deals with competing risks data where the cause of death information is not available or is unreliable. Typically it is used in the analysis of cancer registry data that contains information on survival of cancer patients, who can die either from cancer or from other causes (from here on we shall speak of cancer patients even though this methodology can be used elsewhere as is also shown by one of our examples). The idea is to replace the missing information on cause of death by using the information about general population mortality given in the population mortality tables. All the underlying assumptions related to this issue are described more precisely in the next chapter.

1.1.1 Population mortality tables

Population mortality tables are used to represent the hazard for dying from other causes of cancer patients. They are usually split with respect to age, gender and year of diagnosis and sometimes also with respect to other relevant variables (such as social class or race). We shall refer to these covariates

as the demographic covariates. In this work we use population mortality tables for Slovenia that are split only with respect to the first three variables.

Population mortality tables represent the proportion of people of a certain age and sex that were alive on the 1st of January of any given year and died during that same year. The structure of the available data forces us to assume that the hazard for dying from other causes is a step function that changes twice a year for each patient (once the patient gets one year older and at the beginning of a new year).

Population mortality tables for Slovenia are integrated in the `R` package `relsurv` [24], for other countries they can be obtained from various web pages (e.g. [1]).

1.1.2 Different measures

Everything you can imagine is real.

Pablo Picasso

The main goal of the relative survival analysis is to summarize the survival of cancer patients and compare it to the general (disease free) population or between different populations. Several different measures can be used to this end and they present the information of interest in different ways. Estimators of these measures suffer from the same flaw, i.e. whenever the population mortality tables do not represent the hazard of dying from other causes sufficiently well, they are biased. Therefore the assumption that the population mortality tables represent the hazard of dying from other causes is crucial. When it is violated, the biased estimates are as good as it gets. These estimators can be used anyway but the estimates have to be given a cautious interpretation (taking into account the fact that the hazard of dying from other causes is misrepresented).

The first measure that we shall present is the *expected survival* $S_P(t)$. For a group of cancer patients, it represents the survival of their counterparts, i.e. individuals with the same demographic covariates but without cancer. To obtain this measure, we only have to extract the information about survival from the population mortality tables. Therefore, this measure can be estimated whenever the population mortality tables are available. Whenever we can assume that the population mortality tables are split with respect to all the covariates that crucially effect survival, this measure represents how

the cancer patients would live if they did not have the disease (and hence the name expected survival is justified). This assumption is often violated, e.g. population mortality tables are not split with respect to smoking and therefore such conclusions can not be made for smokers. This measure can be reported on an individual or group level and is specific for the field of relative survival (not part of competing risks data analysis).

The expected survival of a group of patients can be further used to define the *relative survival ratio* $S_R(t)$, which is calculated simply as the ratio between the overall survival $S_O(t)$ (treating all causes of death equally) and the expected survival

$$S_R(t) = \frac{S_O(t)}{S_P(t)}.$$

For any heterogeneous group of patients, this quantity is not a survival function and there is no theoretical reason why it should be nonincreasing or limited above by 1, although it can satisfy both conditions in practice.

The third measure is the *crude probability of death*, which gives the probability that a patient dies from a specific cause up to time t. This is equivalent to the cumulative incidence function in the competing risks setting. The difference stems from the way the two measures are estimated, i.e. population mortality tables are used to represent the hazard of dying from other causes in the estimation of crude probability of death.

Comparisons between different populations with different other cause mortalities (different countries or time intervals of diagnosis) are very common in cancer registry analysis. The goal of such comparisons is to evaluate how cancer is treated in different populations. To make such comparisons, a measure that does not depend on other causes of mortality is needed. The three measures that we have mentioned so far do not satisfy this condition. They have different values when used on populations with the same quality of cancer care but different general mortality. They are not well suited for making comparisons between different populations, therefore a measure called *net survival* has emerged. We shall define it precisely in the next chapter where also all the underlying assumptions shall be presented. This measure does not depend on the hazard of death from other causes. Several different estimators for this measure have been used in the past (and some are still used nowdays) even though Pokhrel and Hakulinen [27] have realized that these approaches may lead to biased estimates when the hazard of death from other causes depends on some common covariates.

An important step forward happened in 2012, when Pohar Perme et al. [26] clearly showed that different estimators that were used to estimate net survival produce biased estimates. They proposed a new consistent estimator of net survival (Pohar Perme estimator, from here on called PP estimator). The next step forward came in 2016 with the paper by Sasieni and Brentnall [31]. As we see it, the main contribution of this paper is in clearly defining the desired properties of the measure of interest, i.e. of a measure that could be used for comparisons between different populations. Instead of optimising the already existing estimators and fighting over their preferability they focused on proposing an alternative measure with desired properties and a way of estimating it.

1.2 Goals of this work

Historically, the field of relative survival analysis has been focused on estimation of measure that will allow comparisons between different populations. The first step towards the solution is net survival because it does not depend on the hazard of dying from other causes. Net survival is also an ideal that is often out of reach in practice as we shall see later on. Practitioners have been trying to estimate it for decades and we would like to warn them about the assumptions, interpretation and issues related to this measure. Our goal is to clearly state what can be estimated without additional assumptions and how it can be interpreted.

In this work we would like to take the same path as in [31] and look at the problem from a distance and thus focus on the measure of interest first. We will propose an alternative (that uses the available information differently compared to the measure defined in [31]) that has a nice interpretation and under additional assumptions equals net survival. We think that net survival is not to be thrown away especially if one is willing to make the assumptions it requires, otherwise the new proposal is the closest we can get. We do not want to advocate the use of net survival, we would rather like to clarify its assumptions and drawbacks.

Research for this work started with the exploration of the properties of the log-rank type test (presented in Chapter 4). During that time we got a deeper understanding of the problems of net survival and started the search for alternatives. We shall first define the measure of interest and its association to the net survival and then explore their assumptions. Afterwards

we shall focus on estimation of the new measure and then we shall return to comparison of net survival curves.

1.3 Notation

'Why' is the only question that bothers people enough to have an entire letter of the alphabet named after it. The alphabet does not go 'A B C D What? When? How?' but it does go 'V W X Why? Z.

Douglas Adams

Here we present the list of symbols used in the text. Although the quantities presented here are not defined yet, the list is included here so that it can be used as a reference list afterwards. We present only the main quantities, subscripts can be used to specify them further. Some of the quantities can also have another meaning in the text and this will be clearly stated when necessary. An estimator for quantity A will be denoted as $\widehat{A}$.

S	survival curve;
λ, Λ	(marginal) hazard and cumulative hazard function; superscript * will be used to denote cause specific quantities;
O, P, E	when used as subscripts they indicate the overall, population (or other cause) and excess terms;
n	sample size, i.e. the number of patients;
i, j, k	indexes counting patients;
t	time;
τ	the longest possible follow up time; time when we want to have estimates or make comparisons;
D_i, Z_i	demographic and other covariates of interest for the ith patient;
X_i	vector of covariates for the ith patient, $X_i = (D_i, Z_i)$;
$T_i(\widetilde{T}_i)$	random variable denoting the ith patient's (observed) time;
δ_i	indicator function for the censoring status (0=censoring, 1=event);
$N_i(t)$	counting process for the ith patient;
$Y_i(t)$	at risk process for the ith patient.

1.4 Simulations and real data

We use simulations and real data through out this work to illustrate certain points, to check the properties of the (proposed) methods and to illustrate their behaviour.

1.4.1 Simulations

Simulations used in this work have some common features. We present them here in order to have a clearer plot and to explain how they reflect reality.

For the ith patient we simulate two latent times T_{Ei} and T_{Pi}, i.e. time to death from cancer and from other causes and then take their minimum to be the observed time ($T_i = \min(T_{Ei}, T_{Pi})$). Therefore, we know the true latent times of death from cancer and we also know what really happens in the hypothetical world where people can only die from cancer. Since this is the world on which the *net survival* measure operates (see Chapter 2 for details), this enables us to assess the properties of the estimators of this measure.

Time T_{Ei} is generated from the Cox model $\lambda(t|X_i) = \lambda_0(t)e^{\beta X_i}$, but the baseline hazard is taken to be constant since we are not primarily interested in its effects (i.e. $\lambda_0(t) = \lambda_0$ is from exponential model). Time T_{Pi} is generated for each individual separately based on the vector of demographic covariates D_i and using Slovene population mortality tables. We generate data under the assumption that both hazards are constant in short intervals, i.e. we split time interval in a number of short intervals and generate individual times using probability integral transform [29, p. 63, 353]. If they fall in a short interval, we add the generated times to the left end of the short interval and proclaim this sum to be the event time. Otherwise, we move to the next short interval and repeat the procedure. We also determine the maximal follow-up time and those that do not have an event up to that time are considered to be censored. In this way we achieve that the only ties in our data are at the end of follow up time.

We allow for external censoring. We independently generate censoring times C_i for all individuals. We generate them from a uniform distribution $U[0, a]$, where we choose a so that we get the desired level of censoring. In Chapter 4 we also use exponential censoring distributions $\text{Exp}(\lambda)$, where λ is also chosen to give the desired amount of censoring.

For each individual we define his observed times as $\widetilde{T}_i = \min(T_i, C_i)$. We also define $\delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise. We shall refer to these data as the *real world* data, as opposed to the *hypothetical world* data $(\widetilde{T}_{Ei}, \delta_{Ei})$, where $\widetilde{T}_{Ei} = \min(T_{Ei}, C_i)$ and $\delta_{Ei} = 1$ if $T_{E,i} \leq C_i$ and 0 otherwise.

We do not attempt to build realistic simulation scenarios, we rather use simple ones where we can track back the observed properties to their real cause. We change only the parameters that are known to have a large effect on measures and estimators in the field of relative survival and keep others simple (for example, a constant baseline excess hazard is of course completely unrealistic, but also irrelevant when we are using non-parametric estimation that makes no assumptions about the baseline).

Covariate distribution and effect sizes are chosen to give certain proportions of events or certain proportion of events of a given cause among all events (since these are the parameters that have the largest impact on the estimators in the field of relative survival). Covariates are generated independently of each other.

Further details are provided within specific sections.

1.4.2 Data

In this work we use two data sets. They are briefly presented below and used mainly to illustrate the methods. The goal of this work is definitely not to draw conclusions, a much more detailed analysis should be performed to this end.

Colon cancer data

We use the data of all Slovene female colon cancer patients - 1538 female patients diagnosed between 1st January 1994 and 30th December 2000, data are taken from Slovene cancer registry [38]. Their age span is between 26 and 85. We will use this data to illustrate some points that one has to be careful about when estimating long term survival. This data set was chosen on purpose - as an example of a truly bizarre behaviour we can get with the routine use of the (PP) estimator.

AMI data

Data comes from a study of survival of patients with acute myocardial infarction (AMI) which was carried out at the University Clinical centre in Ljubljana, Slovenia [35]. We will be mainly interested in the excess mortality that these patients experience due to the infarction, we focus on a subgroup of 494 patients, aged between 45 and 75, who were recruited from 1984 and 1986 and followed for 10 years. The patients were included in the study at the time of release from the hospital after an infarction, the end point was death of any cause, whereas cause-of-death information was not available. Several variables were recorded at the time of admission, but we focus here on sex and age only. In this work we will not be interested in the overall mortality of these patients, but rather in the excess hazard that can be attributed to their cardio-vascular disease and the effect of sex on it.

Chapter 2

The measure of interest

'The story so far:
In the beginning the Universe was created.
This has made a lot of people very angry and been widely regarded as a bad move.'
Douglas Adams

In this chapter we will formally introduce the net survival measure, its underlying assumptions and propose a new measure.

2.1 Additive model

Let T be a continuous random variable denoting the failure time (i.e. time to death) and let C be the consoring time. Let us denote $\widetilde{T} = \min(T, C)$.

A common starting point in the field of relative survival is the competing risks model

$$\lambda_O(t|X) = \lambda_E^*(t|X) + \lambda_P^*(t|D), \tag{2.1}$$

often referred to as the additive model for the overall hazard $\lambda_O(t|X)$ [10]. This model assumes that the causes of death can be split to deaths from cancer and deaths from other causes and that the same holds for hazards, i.e. that the hazard for dying from any cause λ_O can be split into the sum of the cause specific hazards for dying from cancer λ_E^* and from other causes λ_P^* (also referred to as the population hazard).

2.2 Latent times and definition of net survival

Literature on relative survival often starts with the additive model (2.1) although the hazards on the right hand side are not precisely defined. It is often expected from the reader to distinguish and recognize whether the cause specific or marginal hazard was used (as we shall see later the cause specific hazard is used to define the crude probability of death whereas marginal hazard is used to define net survival). We believe this is a common source of misconceptions, therefore we properly define both. This should help any neophyte in this field to get a deeper understanding of the problems and to distinguish between the cause specific and marginal hazard.

'Time is an illusion. Lunchtime doubly so.

Douglas Adams

Let T_E and T_P be two random variables, denoting latent times to death from cancer and from other causes, respectively. Then $T = \min(T_E, T_P)$. Let X represent a vector of covariates that consists of demographic covariates D and other covariates of interest Z, i.e. $X = (D, Z)$ and let δ_E and δ_P denote indicators of death from cancer and from other causes, respectively. Using this notation, we can write the cause specific hazards from the model (2.1) as

$$\begin{aligned}\lambda_E^*(t|X) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t, \delta_E = 1 | T \geq t, X),\\ \lambda_P^*(t|D) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t, \delta_P = 1 | T \geq t, D).\end{aligned} \tag{2.2}$$

On the other hand, the marginal hazards equal

$$\begin{aligned}\lambda_E(t|X) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T_E < t + \Delta t | T_E \geq t, X),\\ \lambda_P(t|D) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T_P < t + \Delta t | T_P \geq t, D).\end{aligned} \tag{2.3}$$

Marginal hazard is thus calculated conditional on one latent time only whereas the cause specific hazard is calculated conditional on the minimum of both latent times T (and as such does not require the existence of latent times).

Generally, these two types of hazard are different and *net survival* is defined using the marginal hazard as

$$S_E(t|X) = \exp\Big(- \int_0^t \lambda_E(u|X)du\Big). \tag{2.4}$$

This measure does not depend on λ_P by definition and is therefore an ideal candidate for comparisons between different populations. However, it represents survival function in the world where only the latent time T_E operates, i.e. in a *hypothetical* world where patients can die only from cancer. This is one of the reasons why this measure has drawn criticism. Issues related to this measure will be further investigated in Section 2.4.

2.3 Marginal relative survival

To avoid the latent time issues of net survival, we propose a new measure. We start from the model

$$\lambda_O(t|X) = \widetilde{\lambda}_E(t|X) + \widetilde{\lambda}_P(t|D). \tag{2.5}$$

This model looks very similar to (2.1) but we make fewer assumptions. In this case the term $\widetilde{\lambda}_P(t|D)$ represents the hazard for death from other causes for patient's counterpart who is defined by the vector of demographic covariates D. The term $\widetilde{\lambda}_P(t|D)$ is obtained from the population mortality tables. If they include all the covariates that have an impact on the other cause mortality, then $\widetilde{\lambda}_P(t|D) = \lambda_P(t|D)$ (this distinction is used to emphasize that this is often not true, e.g. smoking indicator is an important covariate which is typically missing from the population mortality tables). The term $\widetilde{\lambda}_E(t|X)$ represents an additional excess term that each patient gets because of his disease. We are not assuming that $\widetilde{\lambda}_E(t|X)$ is a hazard function nor that it is positive. In this way we get a very general framework which allows both higher and lower overall hazard for patients. However, the price for this generality is that we will not be able to interpret the new measure as a survival function.

Our interest lies in the excess term $\widetilde{\lambda}_E(t|X)$, which is defined as

$$\widetilde{\lambda}_E(t|X) = \lambda_O(t|X) - \widetilde{\lambda}_P(t|D).$$

It is common to have measures on survival scale (and not on the (cumulative) hazard scale) in the field of relative survival analysis. To ensure comparability

we use the functional $f \mapsto \exp(-\int_0^t f(u)du)$ on the last equation to get

$$\widetilde{S}_E(t|X) = \frac{S_O(t|X)}{\widetilde{S}_P(t|D)}. \tag{2.6}$$

This is the new measure on the individual level (for someone with covariates X) - we shall refer to it as *relative survival*. To get a measure on a group level we can simply integrate it over the distribution of X (H denotes the distribution function of X)

$$\widetilde{S}_E(t) = \int \widetilde{S}_E(t|x)dH(x). \tag{2.7}$$

We shall refer to this measure as *marginal relative survival* and we shall interpret it as the average of ratios between individual overall survival and his counterpart's survival. We can also allow comparisons with populations with pregiven distributions of demographic covariates by introducing additional weights w, i.e.

$$\widetilde{S}_E(t) = \int w(x)\widetilde{S}_E(t|x)dH(x). \tag{2.8}$$

Whenever the assumptions of net survival (i.e. Equation (2.10)) are valid and $\widetilde{\lambda}_P(t|D) = \lambda_P(t|D)$ holds, then also $\widetilde{\lambda}_E(t|X) = \lambda_E(t|X)$ holds. This means that marginal relative survival equals net survival and can be interpreted as a survival function. The equality $\widetilde{\lambda}_P(t|D) = \lambda_P(t|D)$ is typically invalid when the population mortality tables do not represent the other cause mortality of the group of patients well, e.g. if the group of patients consists of smokers. Therefore $\widetilde{\lambda}_E$ is a generalization of λ_E. We will use $\widetilde{\lambda}_E$ whenever we will discuss both of them. We will use $\widetilde{\lambda}_P$ in the same manner.

2.4 Underlying assumptions and issues of net and marginal relative survival

2.4.1 Availability of data on excess term

An important drawback lies in the way both net survival and marginal relative survival summarize the excess term information - by definition, the information on $\widetilde{\lambda}_E$ should be available throughout the follow-up time for each individual of the cohort. Subsection 3.3.4 presents an example where

this condition is clearly not met (the data set includes some patients that are very old at the time of diagnosis and we cannot expect that they will give us much information about $\widetilde{\lambda}_E(t|X)$ for large values of t since their overall survival probability is very low). In such a case estimators of both quantities exhibit large variability and peculiarly jumpy behaviour. This has been observed for the PP estimator [9, 33] and has led some authors to search for alternative estimators [18]. As we shall show, the large variability follows from the definition of the measures and is hence not a property of a particular estimator. In particular, measure (2.7) is an integral of ratios (2.6) across the distribution of covariate vector X. In (2.6) the overall survival $S_O(t|X)$ is weighted by the factor $1/\widetilde{S}_P(t|D)$. For some values of the covariate vector D, this factor can get extremely large values when t increases (e.g. for a patient that is very old at the time of diagnosis, the value of his counterpart's long term survival approaches zero and the reciprocal can have extremely large values) and contributions of such ratios prevail in the value of the measure which can therefore have extremely large values for large values of t. When the value of $\widetilde{S}_P(t|D)$ is close to zero, we cannot expect to extract any information about $\widetilde{S}_E(t|X)$. We therefore propose that the measure $\widetilde{S}_E(t|X)$ should be used only for subgroups of patients for whom we might expect that $\widetilde{S}_P(t|\cdot)$ is large enough at all times.

Sasieni and Brentnall [31] instead propose a measure that can again be seen as a weighted average of the excess term, but defines weights in a way that optimizes the use of the information available, i.e. considers the older patients for a period shorter than the whole follow-up time. They propose to use weights of the form $\widetilde{S}^0_P(t|D)/\widetilde{S}_P(t|D)$ where $\widetilde{S}^0_P(t|D)$ represents the population survival for patient's counterpart from a reference population. This reference population is chosen arbitrarily but should have lower population survival than the country of interest. This ensures that the weights $\widetilde{S}^0_P(t|D)/\widetilde{S}_P(t|D)$ are below one and thus prevents that a certain subgroup of D would dominate the measure. The interpretation of the measure obtained is rather complex and the authors themselves propose that, for a non-specialist audience, it should be described merely as a standardized relative survival index. They claim the same is true for net survival [31].

We agree that the weighting proposed by Sasieni and Brentnall [31] uses the available interpretation better and has nicer statistical properties (i.e. smaller variability of long term estimates) but the cost is that their measure is not directly comparable to the previous results. We believe that particularly the marginal relative survival has an easy to grasp interpretation as the average ratio, but agree that net survival is often interpreted overly loosely

without paying the proper respect to all the assumptions it warrants and the comparability between populations is taken for granted. In the following two subsections, we state these assumptions and thus warn against oversimplifications.

2.4.2 Assumptions affecting comparability between populations

Net survival is often referred to as the measure 'that takes into account the population mortality differences' or the measure 'that does not depend on the population mortality hazard', its comparability between groups with different population mortality is hence a key issue.

- In terms of comparability, the key issue is the assumption that λ_P is given by the population mortality tables, i.e. $\lambda_P = \widetilde{\lambda}_P$. This is often not true, since D does not include all the covariates that affect both λ_P and λ_E, for example, population tables do not include smoking even though this would be crucial for lung cancer data.
  With this assumption violated, populations differing with respect to these, missing covariates (e.g. smoking), become incomparable. To see this, say we wish to compare S_E between Slovene and Chinese male lung cancer patients. A vast majority of these patients are smokers and it is clear that the λ_P calculated on both smokers and non-smokers in the population is underestimating their other-cause mortality, hence their λ_E shall be overestimated. However, since the prevalence of smoking in the two countries is very different (approx 20% in Slovenia vs 50% in China), the amount of this bias shall be very different in the two countries, thus making the estimated $\widehat{S}_E$ ($\widehat{\widetilde{S}}_E$) differ considerably even if the true λ_E ($\widetilde{\lambda}_E$) is in fact the same. Several papers have been published investigating this topic (see e.g. [14]).

- Equation
$$\lambda_O(t|X) = \lambda_E(t|X) + \lambda_P(t|D) \tag{2.9}$$
  is commonly stated as the basic starting model when defining net survival, even though it is usually not clearly stated that λ_E and λ_P are used and not λ_E^* and λ_P^*. Note that the overall hazard λ_O is the con-

ditional probability of dying:

$$\lambda_O(t|X) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t | T \geq t, X)$$

whereas in order for S_P and S_E to be survival functions, λ_P and λ_E should be defined as in (2.3), hence implicitly assuming the existence of latent times T_E and T_P.
Equation (2.9) holds only if we assume that equality (2.10) holds (the assumption for λ_P is analagous):

$$\begin{aligned} \lambda_E^*(t|X) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T_E \leq t + \Delta t | T_P \geq t, T_E \geq t, X) \\ &= \lambda_E(t|X) \end{aligned} \tag{2.10}$$

A sufficient but not necessary assumption would be that T_P and T_E are independent conditional on X. The assumption (2.10) is necessary if S_E is to be independent of T_P.

- The marginal relative survival of course depends on the distribution of covariates X, a difference in $\widetilde{S}_E$ between two populations might therefore stem from a different covariate distribution H. To take this into account, the formula (2.8) can be used as it allows weights w. The idea of the weights is to report results with a pregiven standard distribution of covariates (see e.g. [6] for age-standardization). While this is commonly done (see e.g.[8]), such a standardization is of course only crude, and setting weights that ensure truly comparable results may not be easy. To understand the observed differences in $\widehat{\widetilde{S}}_E$ one shall usually rather resort to regression models.

2.4.3 Assumptions affecting interpretation

- When the population mortality tables do not adequately describe λ_P, the interpretation of net survival fails. However, this assumption is not needed if $\widetilde{S}_E$ is interpreted as the average ratio - the observed survival of an individual is compared to his counterparts in the population, where the counterparts are defined by the covariates D that are available (the patient does not need to be equal to the defined population in all aspects but the fact he has been diagnosed with the disease). The comparability between populations is of course lost regardless of the interpretation.

- The assumption (2.10) cannot be verified from the observed data [17, p. 259-261], in fact verifying it could only make sense if the T_E and T_P did not indicate the times to death, but rather the times to some other events which do not preclude one another. Without this assumption, net survival is not identifiable with real world data (see e.g. [37, 3]), as the next example shows. Say the data follow the additive model in equation (2.9) with T_E and T_P conditionally independent, the observed times T are calculated as the minimum of the two latent times. Using the times T_E only, we estimate survival in the hypothetical world using Kaplan-Meier. Using the observed time T we add the curves estimated by PP estimator and the new approach to the plot (see Figure 2.1). We shall refer to this data set as data set A. There are various scenarios of dependency between T_P and T_E which would result in the same observed times. To get some, we changed the time of the other event after the first one happened. Two extreme cases in terms of the marginal distribution of T_E can be imagined: after an individual experienced the 'other cause' event ($T = T_P$), the individuals 'dies of cancer' immediately ($T_E = T_P + \Delta t$, dataset B) or, the individual becomes immortal with respect to cancer ($T_E = \infty$, dataset C). Using times T_E only, we again estimate survival in the hypothetical world. The estimators on the three data sets (A, B, C) in Figure 2.1 are wildly different and the value of net survival between the two extrema (B and C) is unidentifiable.

- We believe the historical interest in net survival does not stem from the interest in what would happen if we could remove other causes, rather, it is driven by the need of a measure independent of the population mortality. To avoid the long-standing debate on the sensibility of net survival and the need for untestable assumptions, the more cautious interpretation of the marginal relative survival in terms of the average ratio can be used. Without assuming (2.10) (and conditional on λ_P being correct), the extent of comparability has been summarized as the criterium A2 in Sasieni's work [31]: $\widetilde{\lambda}_E$ depends on all causes of mortality, however this dependence is not driven by marginal size of $\widetilde{\lambda}_P$ (as for example with relative survival ratio or cumulative incidence function), but rather of the way the different causes of death are dependent in their joint probability.

All things being said, the marginal relative survival shall in practice never be completely comparable, but it is (possibly with some additional weighting) a

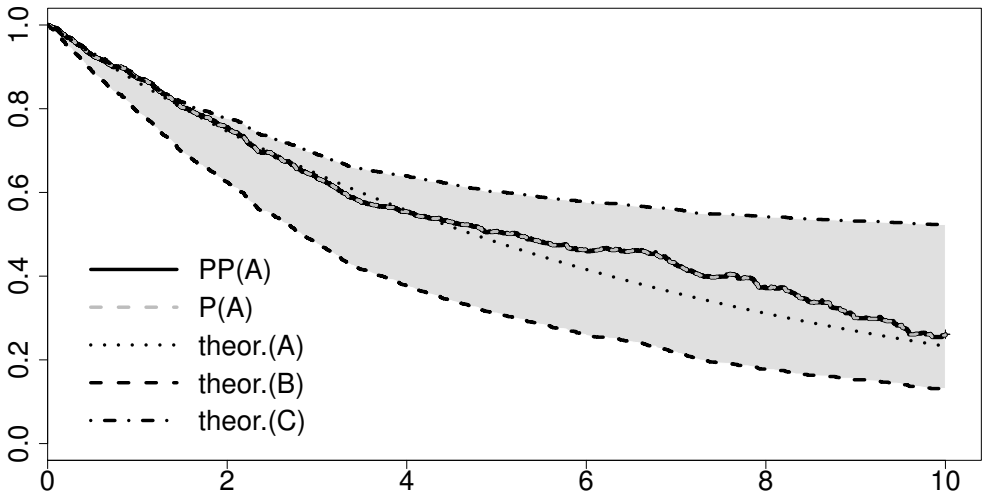

Figure 2.1: Identifiability of net survival: PP is the PP estimator on observed data, P is the pseudo based approach to estimation of the marginal relative survival on observed data, theor. is the Kaplan-Meier estimator on the hypothetical world data

reasonable and simple survival measure to report when focusing on $\widetilde{\lambda}_E$. While the data are never perfect and the above listed assumptions clearly indicate what is relevant when judging them, it is important that the measure at least avoids depending on $\widetilde{\lambda}_P$ directly when this is not necessary (and depends on $\widetilde{\lambda}_P$ only through $\widetilde{\lambda}_E$).

Chapter 3

Estimation of marginal relative survival

In this chapter we shall focus on estimation of the measure defined in (2.7). Whenever we can obtain estimates of the measure on the individual level, we can estimate it for the sample of size n as

$$\widehat{\widetilde{S}}_E(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{S}_{Oi}(t)}{\widetilde{S}_{Pi}(t)}. \tag{3.1}$$

We use notation $\widehat{S}_{Oi}(t) = \widehat{S}_O(t|X_i)$ and similarly for $\widetilde{S}_{Pi}(t)$. We obtain information about $\widetilde{S}_{Pi}(t)$ from the population mortality tables and the only question is how to estimate individual overall survival $\widehat{S}_{Oi}(t)$.

We will first consider a special case of complete data, i.e. data without censored observations.

Data without censored observations We assume that all individuals have the same potential follow up time τ and that for each patient his censoring time is $C_i = \tau$. We can use indicator function $I(\widetilde{T}_i > t)$ to estimate individual overall survival. This indicator equals 1 up to the ith individual's event time and 0 afterwards. For those that are censored at the end of follow up, the indicator function remains at 1 for the entire observation period. Figure 3.1 presents an example of two indicator functions for two individuals with and without an event.

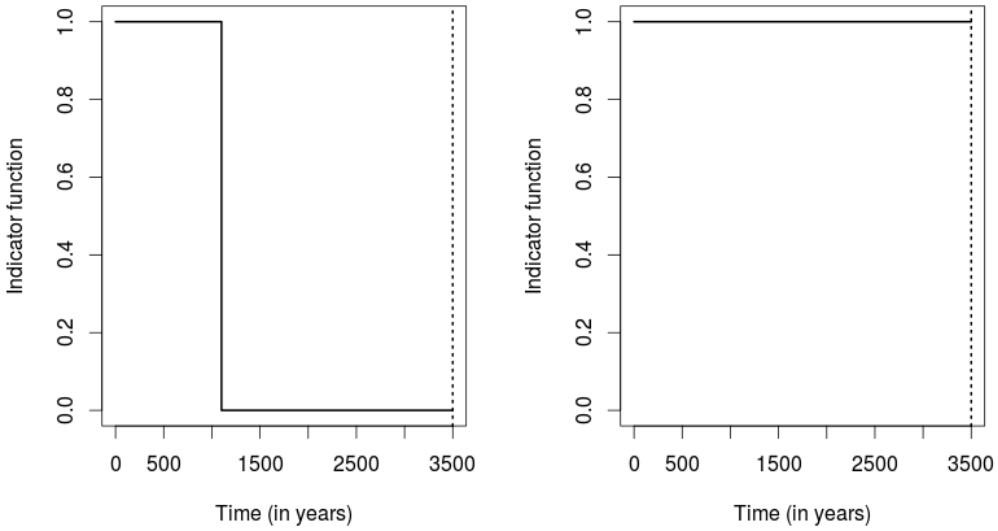

Figure 3.1: Indicator functions, left: patient with an event, right: patient, censored at τ

We can simply use $I(\widetilde{T}_i > t)$ as $\widehat{S}_{Oi}(t)$ in (3.1) and obtain a nonparametric estimator of (2.7). This allows us to obtain estimates up to time τ (if some of the patients are censored at that time, the estimate will be nonzero, otherwise the estimator can reach zero at an earlier time).

Data with censored observations In this case some of the patients may get censored before τ. The idea presented above does not work because we can not use indicator functions to estimate survival of those censored after they are censored. Instead, we can use pseudo observations [4] as estimators of individual overall survival.

3.1 Pseudo observations

Suppose we are interested in $f(T)$, where f is an arbitrary function. Let $\theta = E(f(T))$. With complete data, the survival times T_i are observed for all individuals and θ can be simply estimated as $\frac{1}{n}\sum_{i=1}^{n} f(T_i)$. Assume further that whenever some observations are censored (and not all $f(T_i)$ are

observed) there exists a well behaved estimator $\widehat{\theta}$ for $\theta = E(f(T))$. In this case the pseudo observation $\widehat{\theta}_i$ for $f(T_i)$ for ith individual is defined as

$$\widehat{\theta}_i = n \cdot \widehat{\theta} - (n-1) \cdot \widehat{\theta}^{(-i)}, \tag{3.2}$$

where $\widehat{\theta}^{(-i)}$ is the estimator applied to the sample of size $n-1$ obtained from the original sample by eliminating the ith observation.

Pseudo observations can be calculated for several different measures θ, a thorough explanation for some of them can be found in [5]. We present two examples just to illustrate the generality of this approach.

Survival probability Let t be a fixed time and $f(T) = I(T > t)$. In this case $E(f(T)) = P(T > t) = S(t) = \theta$. Kaplan-Meier estimator $\widehat{S}(t)$ can be used to obtain the estimate of $S(t)$ on the whole sample. The ith pseudo observation is

$$\widehat{S}_i(t) = n \cdot \widehat{S}(t) - (n-1) \cdot \widehat{S}^{(-i)}(t). \tag{3.3}$$

Since the Kaplan-Meier estimator changes only at event times, the same holds for the pseudo observations defined in (3.3).

Restricted mean survival time The t-restricted mean survival time is defined as $\mu_t = E(T \wedge t)$, i.e. $f(T) = T \wedge t$. Since

$$T \wedge t = \int_0^{T \wedge t} ds = \int_0^t I(T > s) ds,$$

we have $\mu_t = E(T \wedge t) = \int_0^t S(s) ds$. A natural estimator of μ_t is thus $\widehat{\mu}_t = \int_0^t \widehat{S}(s) ds$. The ith pseudo observation is then

$$\widehat{\mu}_{t,i} = n \cdot \int_0^t \widehat{S}(s) ds - (n-1) \cdot \int_0^t \widehat{S}^{(-i)}(s) ds = \int_0^t \widehat{S}_i(s) ds$$

simply expressed with (3.3).

Pseudo observations can also be used in the competing risk setting for cumulative incidence functions and years lost or for state occupation probabilities in the general multistate models. They thus present a general tool that has many different applications and allows us to obtain individual estimates (even for censored individuals) at all times (up to the largest observed time). Regardless of the application, they have to be used for all individuals and not just for the censored ones. We will focus on pseudo observations for survival probability since they are the backbone of the new estimation approach.

3.1.1 Properties of pseudo observations for survival probability

Pseudo observations for complete data When the data set does not include censored observations, i.e. $\widetilde{T}_i = T_i \wedge \tau$ for $i = 1 \ldots, n$, the Kaplan Meier estimator simplifies to the proportion of patients still being at risk $\widehat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{T}_i > t)$. Similarly, the ith pseudo observation simplifies to

$$\widehat{S}_i(t) = n \cdot \left(\frac{1}{n}\sum_{j=1}^{n} I(\widetilde{T}_j > t)\right) - (n-1) \cdot \left(\frac{1}{n-1}\sum_{j\neq i} I(\widetilde{T}_j > t)\right) = I(\widetilde{T}_i > t).$$

Interpretation of pseudo observations Figure 3.2 presents plots for two pseudo observations, the left one is for a patient with an event and the right one is for a censored patient. The left one is close to 1 while the patient is still at risk, i.e. still alive and becomes close to 0 after he has an event, it looks similar to the indicator function for someone with an event (see Figure 3.1). Pseudo observation for a censored individual is close to 1 while he is still at risk (as would be an indicator function for a censored individual) but is defined also after he gets censored (in contrast to the indicator function). After the censoring time, the pseudo observation slowly decreases meaning that the estimated probability of having had an event is increasing. They thus have intuitively sensible behaviour.

Cautious observer can also notice that pseudo observations can have values below zero or above one (see Figure 3.2) but this is not a problem since they are only estimating probability (which should be between zero and one). Furthermore, Stute and Wang showed [36] that the average of pseudo observations (calculated using the Kaplan-Meier estimator) $\frac{1}{n}\sum_{i=1}^{n} \widehat{S}_i(t)$ equals the Kaplan-Meier estimator with the only two possible exceptions being the largest two observed times (the difference appears only when the largest observed time belongs to an event and the one before that belongs to censoring). This means that the ith pseudo observation can be interpreted as the contribution of the ith patient to the estimate of overall survival obtained by using the Kaplan-Meier estimator.

Their result also explains why pseudo observations have to be used for all individuals and not only for the censored ones. If we use pseudo observations only for the censored individuals and the indicator functions for the rest, the average of these individual values would not equal the Kaplan-Meier

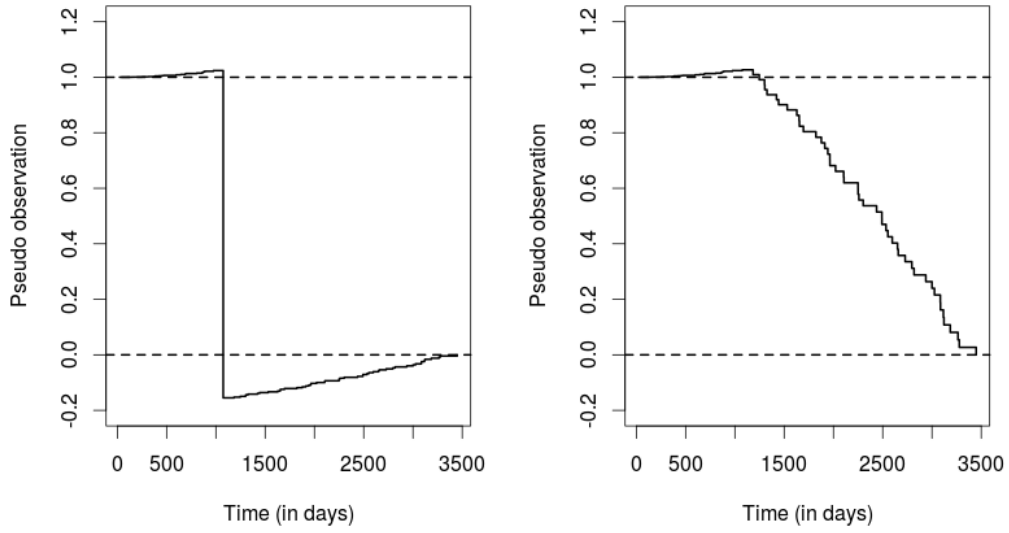

Figure 3.2: Pseudo observations for survival probability, left: patient with an event, right: censored patient

estimate. Since the Kaplan-Meier estimator is consistent, the average of individual values would be biased (at least on large samples).

Asymptotic behaviour of pseudo observations Graw et al. [15] explored the asymptotic properties of pseudo observations for cumulative incidence function in the competing risks setting, a special case being the situation where we do not distinguish between causes of death (this holds when we are considering overall survival). Let $\widehat{F}$ denote the Aalen-Johansen estimator of cumulative distribution function for all causes combined (see [2, p. 288]) and $\widehat{F}_i$ the ith pseudo observation for the same quantity. They used von Mises expansion ($\dot{\psi}(\cdot)$ denotes the first order influence function) to show that for each i

$$\widehat{F}_i(t) = \dot{\psi}(\widetilde{T}_i, \delta_i, X_i) + F(t) + o_P(1), \tag{3.4}$$

and

$$E\big(\widehat{F}_i(t)\big) = F(t) + o_P(1), \tag{3.5}$$

see Lemmas 1 and 2 from [15] for details. This is true under some regularity conditions, the most important of them requiring that the censoring times C_i are independent of (T_i, X_i). We can prove (using mathematical induction)

that $\widehat{S}(t) = 1 - \widehat{F}(t)$, where $\widehat{S}(t)$ is the Kaplan-Meier estimator of survival. Then $\widehat{F}_i(t) = n \cdot \widehat{F}(t) - (n-1) \cdot \widehat{F}^{(-i)}(t) = n \cdot \big(1 - \widehat{S}(t)\big) - (n-1) \cdot \big(1 - \widehat{S}^{(-i)}(t)\big) = 1 - \widehat{S}_i(t)$. Using this relation in (3.4) and (3.5) together with the linearity of expectation, we can establish that

$$\widehat{S}_i(t) = \dot{\psi}(\widetilde{T}_i, \delta_i, X_i) + S(t) + o_P(1) \tag{3.6}$$

and

$$E\big(\widehat{S}_i(t)\big) = S(t) + o_P(1). \tag{3.7}$$

Equation (3.6) shows that the ith pseudo observation asymptotically depends solely on the observed information of the ith patient. When the remainder terms can be neglected, pseudo observations are independent. This is trivially true also for finite samples when the data set does not include censored observations because the pseudo observations in this case simplify to indicator functions. Jacobsen and Martinussen [16] showed that when using pseudo observations as response variables in models and estimating the variance of the coefficients estimated using the generalized estimating equations the remainder term generally can not be neglected. They showed that the second order approximation has to be used in this case and used simulations to explore how well both approaches perform in practice. Their simulation results indicate that the second-order term only plays a role when there is a strong correlation between the covariate and the survival time together with a high-censoring percentage [16].

We shall assume that pseudo observations are independent for the derivation of approximative formula for the variance of the new estimator. We shall refer to this property as approximative independence of pseudo observations. We will also present another (precise) approach to variance estimation and comparison of their results will shed some light on this assumption.

Equation (3.7) shows that the expected value of each pseudo observation asymptotically converges to the survival probability, i.e. to the quantity that is being estimated for each patient by the pseudo observation.

3.2 Variance derivation and estimation

Someone told me that each equation I included in the book would halve the sales.
Stephen Hawking

We propose to estimate marginal relative survival as

$$\widehat{\widetilde{S}}_E(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{S}_{Oi}(t)}{\widetilde{S}_{Pi}(t)},$$

where $\widehat{S}_{Oi}(t)$ are pseudo observations for survival probability calculated with Kaplan-Meier estimator. We shall call it the pseudo based estimator and denote it by P.

3.2.1 Variance derivation

We derive the variance formula for the estimator $\widehat{\widetilde{S}}_E(t)$. We treat the values $\widetilde{S}_{Pi}(t)$ as constants (this is a common approach in the field of relative survival analysis).

We first rewrite the estimator $\widehat{\widetilde{S}}_E(t)$ using the definition of pseudo observation for survival probability (3.3). We get

$$\widehat{\widetilde{S}}_E(t) = \widehat{S}_O(t)\cdot\sum_{i=1}^{n}\frac{1}{\widetilde{S}_{Pi}(t)} - \frac{n-1}{n}\cdot\sum_{i=1}^{n}\frac{\widehat{S}_O^{(-i)}(t)}{\widetilde{S}_{Pi}(t)}. \tag{3.8}$$

Variance equals

$$\begin{aligned}
\operatorname{var}\left(\widehat{\widetilde{S}}_E(t)\right) &= \left(\sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}(t)}\right)^2 \cdot \operatorname{var}\left(\widehat{S}_O(t)\right) \\
&\quad - 2 \cdot \frac{n-1}{n} \cdot \sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}(t)} \cdot \operatorname{cov}\left(\widehat{S}_O(t), \sum_{j=1}^{n} \frac{\widehat{S}_O^{(-j)}(t)}{\widetilde{S}_{Pj}(t)}\right) \\
&\quad + \left(\frac{n-1}{n}\right)^2 \cdot \operatorname{var}\left(\sum_{i=1}^{n} \frac{\widehat{S}_O^{(-i)}(t)}{\widetilde{S}_{Pi}(t)}\right) \\
&= \left(\sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}(t)}\right)^2 \cdot \operatorname{var}\left(\widehat{S}_O(t)\right) \\
&\quad - 2 \cdot \frac{n-1}{n} \cdot \sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}(t)} \cdot \left\{ \sum_{j=1}^{n} \frac{1}{\widetilde{S}_{Pj}(t)} \cdot \operatorname{cov}\left(\widehat{S}_O(t), \widehat{S}_O^{(-j)}(t)\right)\right\} \\
&\quad + \left(\frac{n-1}{n}\right)^2 \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\widetilde{S}_{Pi}(t)\widetilde{S}_{Pj}(t)} \cdot \operatorname{cov}\left(\widehat{S}_O^{(-i)}(t), \widehat{S}_O^{(-j)}(t)\right).
\end{aligned} \tag{3.9}$$

With the assumption of approximative independence of pseudo-observations (defined at the end of Subsection 3.1.1), the variance of the estimator $\widehat{\widetilde{S}}_E(t)$ may be approximated by

$$\operatorname{var}\left(\widehat{\widetilde{S}}_E(t)\right) = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{S}_{Oi}(t)}{\widetilde{S}_{Pi}(t)}\right) \approx \frac{1}{n^2}\sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}^2(t)} \operatorname{var}\left(\widehat{S}_{Oi}(t)\right) \tag{3.10}$$

Variance of $\widehat{S}_{Oi}(t)$ can be expressed as

$$n^2 \operatorname{var}\left(\widehat{S}_O(t)\right) - 2n(n-1)\operatorname{cov}\left(\widehat{S}_O(t), \widehat{S}_O^{(-i)}(t)\right) + (n-1)^2 \operatorname{var}\left(\widehat{S}_O^{(-i)}(t)\right)$$

by using the definition of pseudo observation for survival probability (3.3). This expression can be used to replace $\operatorname{var}\left(\widehat{S}_{Oi}(t)\right)$ in (3.10). We obtain the following approximative formula for the variance of $\widehat{\widetilde{S}}_E(t)$

$$\begin{aligned}
\operatorname{var}\left(\widehat{\widetilde{S}}_E(t)\right) &\approx \operatorname{var}\left(\widehat{S}_O(t)\right) \cdot \left(\sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}^2(t)}\right) \\
&\quad - 2 \cdot \frac{n-1}{n} \cdot \sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}^2(t)} \cdot \operatorname{cov}\left(\widehat{S}_O(t), \widehat{S}_O^{(-i)}(t)\right) \\
&\quad + \left(\frac{n-1}{n}\right)^2 \sum_{i=1}^{n} \frac{1}{\widetilde{S}_{Pi}^2(t)} \cdot \operatorname{var}\left(\widehat{S}_O^{(-i)}(t)\right)
\end{aligned} \tag{3.11}$$

Comparison of formulas (3.9) and (3.11) reveals that we have to calculate only n different variances $\mathrm{var}\big(\widehat{S}_O^{(-i)}(t)\big)$ when using formula (3.11) whereas we also have to calculate all $\binom{n}{2}$ different covariances $\mathrm{cov}\big(\widehat{S}_O^{(-i)}(t), \widehat{S}_O^{(-j)}(t)\big)$ to use formula (3.9). Estimator based on the approximative formula (3.11) is therefore much less computationally demanding and will work much faster on large samples.

3.2.2 Variance estimation

To estimate the variance following the formulas (3.9) and (3.11), we have to estimate the variance $\mathrm{var}\big(\widehat{S}_O^{(-i)}(t)\big)$ and the following two covariances: $\mathrm{cov}\big(\widehat{S}_O(t), \widehat{S}_O^{(-i)}(t)\big)$ and $\mathrm{cov}\big(\widehat{S}_O^{(-i)}(t), \widehat{S}_O^{(-j)}(t)\big)$. To this end, we shall build upon ideas for the estimation of variance of the Kaplan-Meier estimator.

Let $N_i(t) = I(\widetilde{T}_i \leq t, \delta_i = 1)$, $Y_i(t) = I(\widetilde{T}_i \geq t)$, $N(t) = \sum_{i=1}^n N_i(t)$, $N^{(-j)}(t) = \sum_{i \neq j} N_i(t)$ and similarly for Y, $Y^{(-j)}$ and $Y^{(-j,-k)}$. Let $S_O^*(t) = \exp\big(-\Lambda_O^*(t)\big)$, where $\Lambda_O^*(t) = \int_0^t I\big(Y(u) > 0\big)\lambda_O(u)du$ and similarly for $S_O^{*(-i)}(t)$. Remember also that $\widehat{S}_O(t) = \prod_{s \leq t}\big(1 - \Delta\widehat{\Lambda}(s)\big) = \prod_{s \leq t}\big(1 - \frac{\Delta N(s)}{Y(s)}\big)$ and similarly for the same quantities with superscripts.

The terms $\mathrm{var}\big(\hat{S}_O(t)\big)$ and $\mathrm{var}\big(\hat{S}_O^{(-i)}(t)\big)$ are simply the variances of the Kaplan-Meier estimator calculated on the whole and on the reduced samples, respectively. To estimate these two terms, the standard estimators based on counting processes can be used [2]. The remaining covariances can be approximated using the facts that

$$\mathrm{cov}\Big(\frac{\hat{S}_O}{S_O^*} - 1, \frac{\hat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1\Big)(t) = E\Big\langle \frac{\hat{S}_O}{S_O^*} - 1, \frac{\hat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1\Big\rangle(t)$$

and

$$\mathrm{cov}\Big(\frac{\hat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1, \frac{\hat{S}_O^{(-j)}}{S_O^{*(-j)}} - 1\Big)(t) = E\Big\langle \frac{\hat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1, \frac{\hat{S}_O^{(-j)}}{S_O^{*(-j)}} - 1\Big\rangle(t).$$

These predictable covariation processes can be calculated using similar techniques as in [2]. It follows from the Duhamel's equation

$$\frac{\widehat{S}_O(t)}{S_O^*(t)} - 1 = -\int_0^t \frac{\widehat{S}_O(s-)}{S_O^*(s)} \frac{I(Y(s) > 0)}{Y(s)} dM(s)$$

that

$$\left\langle \frac{\widehat{S}_O}{S_O^*} - 1, \frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1 \right\rangle(t) = \left\langle 1 - \frac{\widehat{S}_O}{S_O^*}, 1 - \frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}} \right\rangle(t) =$$

$$\left\langle \int_0^t \frac{\widehat{S}_O(s-)}{S_O^*(s)} \cdot \frac{I\big(Y(s) > 0\big)}{Y(s)} dM(s), \int_0^t \frac{\widehat{S}_O^{(-i)}(s-)}{S_O^{*(-i)}(s)} \cdot \frac{I\big(Y^{(-i)}(s) > 0\big)}{Y^{(-i)}(s)} dM^{(-i)}(s) \right\rangle =$$

$$\int_0^t \frac{\widehat{S}_O(s-)}{S_O^*(s)} \cdot \frac{I\big(Y(s) > 0\big)}{Y(s)} \cdot \frac{\widehat{S}_O^{(-i)}(s-)}{S_O^{*(-i)}(s)} \cdot \frac{I\big(Y^{(-i)}(s) > 0\big)}{Y^{(-i)}(s)} d\langle M, M^{(-i)} \rangle(s).$$

Since

$$d\langle M, M^{(-i)} \rangle(s) = d\langle M^{(-i)} + M_i, M^{(-i)} \rangle(s) = d\langle M^{(-i)} \rangle(s) =$$

$$(1 - \Delta\Lambda(s)) \cdot Y^{(-i)}(s) d\Lambda(s),$$

we obtain the following result

$$\begin{aligned} &\left\langle \frac{\widehat{S}_O}{S_O^*} - 1, \frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1 \right\rangle(t) = \\ &= \int_0^t \frac{\widehat{S}_O(s-)}{S_O^*(s)} \cdot \frac{I\big(Y(s) > 0\big)}{Y(s)} \cdot \frac{\widehat{S}_O^{(-i)}(s-)}{S_O^{*(-i)}(s)} \cdot \frac{I\big(Y^{(-i)}(s) > 0\big)}{Y^{(-i)}(s)} \times \\ &\times \big(1 - \Delta\Lambda_O(s)\big) Y^{(-i)}(s) d\Lambda_O(s) \end{aligned} \tag{3.12}$$

Similar arguments can be used to show that

$$\begin{aligned} &\left\langle \frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}} - 1, \frac{\widehat{S}_O^{(-j)}}{S_O^{*(-j)}} - 1 \right\rangle(t) = \\ &= \int_0^t \frac{\widehat{S}_O^{(-i)}(s-)}{S_O^{*(-i)}(s)} \cdot \frac{I\big(Y^{(-i)}(s) > 0\big)}{Y^{(-i)}(s)} \cdot \frac{\widehat{S}_O^{(-j)}(s-)}{S_O^{*(-j)}(s)} \cdot \frac{I\big(Y^{(-j)}(s) > 0\big)}{Y^{(-j)}(s)} \times \\ &\times \big(1 - \Delta\Lambda_O(s)\big) Y^{(-i,-j)}(s) d\Lambda_O(s). \end{aligned} \tag{3.13}$$

Replacing $S_O^*(s)$ by $\widehat{S}_O(s)$, $S_O^{*(-i)}(s)$ by $\widehat{S}_O^{(-i)}(s)$ and $\Lambda_O(s)$ by $\widehat{\Lambda}_O(s)$ in the expressions (3.12) and (3.13), we arrive at the following estimators of the covariances

$$\widehat{\mathrm{cov}}\left(\frac{\widehat{S}_O}{S_O^*}, \frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}} \right)(t) = \int_0^t \frac{I\big(Y^{(-i)}(s) > 0\big)}{Y^{(-i)}(s) - \Delta N^{(-i)}(s)} \cdot \frac{Y^{(-i)}(s)}{Y^2(s)} dN(s) \tag{3.14}$$

and

$$\begin{aligned}
&\widehat{\mathrm{cov}}\left(\frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}}, \frac{\widehat{S}_O^{(-j)}}{S_O^{*(-j)}}\right)(t) = \\
&= \int_0^t \frac{I\big(Y^{(-i)}(s) > 0\big)}{Y^{(-i)}(s) - \Delta N^{(-i)}(s)} \cdot \frac{I\big(Y^{(-j)}(s) > 0\big)}{Y^{(-j)}(s) - \Delta N^{(-j)}(s)} \times \\
&\times \big(Y(s) - \Delta N(s)\big) \cdot \frac{Y^{(-i,-j)}(s)}{Y^2(s)} dN(s)
\end{aligned} \tag{3.15}$$

Using these two estimators, we can estimate the covariances required in formulas (3.9) and (3.11) by

$$\widehat{\mathrm{cov}}\big(\widehat{S}_O, \widehat{S}_O^{(-i)}\big)(t) = \widehat{S}_O(t) \cdot \widehat{S}_O^{(-i)}(t) \cdot \widehat{\mathrm{cov}}\left(\frac{\widehat{S}_O}{S_O^*}, \frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}}\right)(t)$$

and

$$\widehat{\mathrm{cov}}\big(\widehat{S}_O^{(-i)}, \widehat{S}_O^{(-j)}\big)(t) = \widehat{S}_O^{(-i)}(t) \cdot \widehat{S}_O^{(-j)}(t) \cdot \widehat{\mathrm{cov}}\left(\frac{\widehat{S}_O^{(-i)}}{S_O^{*(-i)}}, \frac{\widehat{S}_O^{(-j)}}{S_O^{*(-j)}}\right)(t).$$

3.3 Comparison of the PP estimator and the pseudo based approach

The work on estimation has initially started because the PP estimator requires numerical integration that hinders theoretical extensions. Furthermore, its formula is quite sophisticated and therefore its properties are difficult to understand.

Some authors noticed very large variability and peculiarly jumpy behaviour with the practical use of the PP estimator [9, 33]. They attributed this to the PP estimator even though we shall illustrate that this is inherent to net survival (or marginal relative survival). This has led some of them to search for alternative estimators [18].

In this section we compare the PP estimator and the pseudo based approach to estimation in theory and both on simulated and real data and discuss all the aspects mentioned above.

3.3.1 The PP estimator

The PP estimator [26] equals

$$\widehat{\Lambda}_E(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)}, \tag{3.16}$$

where $N_i^w(t) = N_i(t)/S_{Pi}(t)$, $Y_i^w(t) = Y_i(t)/S_{Pi}(t)$, $N^w(t) = \sum N_i^w(t)$ and $Y^w(t) = \sum Y_i^w(t)$. This formula uses the inverse probability weighting to account for those that die from other reasons. In particular, the weighting of Y_i and N_i increases the number of patients still at risk and number of events to account for the expected proportion of patients that may have been lost due to the population hazard [26]. The first integral is a weighted version of the Nelson-Aalen estimator of the overall cumulative hazard and the second one represents the average population cumulative hazard.

The first integral in (3.16) is with respect to a counting process and can be represented as a sum. This is not true for the second integral where the integration is with respect to $d\Lambda_{Pi}(u)$ which changes continuously in time. To evaluate the estimator, the second integral has to be numerically approximated. This is a tedious job which has been done in R package relsurv[24] but not in some other implementations of this estimator. Such poor implementations can affect the quality of the estimates.

3.3.2 Comparison in theory

In Subsection 3.1.1 we showed that with complete data (i.e. whenever $C_i = \tau$ for $i = 1, \ldots, n$) $\widehat{S}_{Oi}(t) = I(\widetilde{T} > t)$ and therefore the estimator (3.1) simplifies to

$$\widehat{\widetilde{S}}_E(t) = \sum_{i=1}^n \frac{I(\widetilde{T} > t)}{\widetilde{S}_{Pi}(t)}.$$

When $\widetilde{\lambda}_E = \lambda_E$ and $\widetilde{\lambda}_P = \lambda_P$ this should also hold for any estimator of net survival. The PP estimator is no exception but its formula does not simplify considerably on complete data.

Both the PP estimator and the pseudo based estimator change with a jump at event times and continuously between them. We thus compare the changes of these two estimators at event times and between them.

We shall assume continuously measured time (i.e. only one event can happen at a time).

Let us first consider what happens at event time t. The jump of the estimator (3.1) at that point can be expressed as:

$$\frac{\widehat{\widetilde{S}}_E(t)}{\widehat{\widetilde{S}}_E(t-)} = \frac{\sum_{i=1}^n \frac{I(\widetilde{T}_i>t)}{S_{Pi}(t)}}{\sum_{i=1}^n \frac{I(\widetilde{T}_i\geq t)}{S_{Pi}(t)}} = \frac{\sum_{i=1}^n Y_i^w(t+)}{\sum_{i=1}^n Y_i^w(t)}.$$

On the other hand, the increment of $d\widehat{\Lambda}_E$ in PP estimator at time of event equals

$$d\widehat{\Lambda}_E(t) = \frac{\sum_{i=1}^n dN_i^w(t)}{\sum_{i=1}^n Y_i^w(t)}.$$

Therefore, we can write

$$1-d\widehat{\Lambda}_E(t) = 1-\frac{\sum_{i=1}^n dN_i^w(t)}{\sum_{i=1}^n Y_i^w(t)} = \frac{\sum_{i=1}^n \frac{I(\widetilde{T}_i\geq t)}{S_{Pi}(t)} - \sum_{i=1}^n \frac{I(\widetilde{T}_i=t)}{S_{Pi}(t)}}{\sum_{i=1}^n Y_i^w(t)} = \frac{\sum_{i=1}^n Y_i^w(t+)}{\sum_{i=1}^n Y_i^w(t)}.$$

Consider a short interval of time $[t, t+dt]$ without any events or censorings. In this case, the jump of the estimator (3.1) equals

$$\frac{\widehat{\widetilde{S}}_E(t+dt)}{\widehat{\widetilde{S}}_E(t)} = \frac{\sum_{i=1}^n \frac{I(\widetilde{T}_i>t+dt)}{S_{Pi}(t+dt)}}{\sum_{i=1}^n \frac{I(\widetilde{T}_i>t)}{S_{Pi}(t)}} = \frac{\sum_{i=1}^n Y_i^w(t+dt)}{\sum_{i=1}^n Y_i^w(t+)}.$$

With the PP estimator, the increment can be heuristically written as

$$\begin{aligned} 1 - d\widehat{\Lambda}_E(t) &\approx 1 + \frac{\sum_{i=1}^n Y_i^w(t)\lambda_P(t)dt}{\sum_{i=1}^n Y_i^w(t)} = \frac{\sum_{i=1}^n \frac{Y_i(t)}{S_{Pi}(t)}(1+\lambda_P(t)dt)}{\sum_{i=1}^n Y_i^w(t)} \\ &\approx \frac{\sum_{i=1}^n \frac{Y_i(t)}{S_{Pi}(t)}e^{\lambda_P(t)dt}}{\sum_{i=1}^n Y_i^w(t)} = \frac{\sum_{i=1}^n \frac{Y_i(t)}{S_{Pi}(t+dt)}}{\sum_{i=1}^n Y_i^w(t)} \\ &= \frac{\sum_{i=1}^n Y_i^w(t+dt)}{\sum_{i=1}^n Y_i^w(t)} = \frac{\sum_{i=1}^n Y_i^w(t+dt)}{\sum_{i=1}^n Y_i^w(t+)}, \end{aligned}$$

since $Y_i^w(\cdot)$ are constant over this short time interval.

We have thus heuristically shown that the increment of PP estimator approximately equals the increment of (3.1) in both cases.

3.3.3 Simulation study

Marginal relative survival is a generalization of net survival. We are thus interested in comparison of both estimators when the two quantities are equal. We use simulations to compare them when the usage of both of them is sensible. We simulate latent times T_E and T_P and compare how the estimators perform on observed data. Furthermore, we explore the properties of both formulas for the estimation of variance of the new proposal. We also add results for the Kaplan-Meier method used in the hypothetical world (h) of latent time T_E. These results are added mostly as a check that simulations were properly conducted and as an ideal to which the estimators on observed data should approach.

The key properties we are changing in simulation scenarios are the sample size and the proportion of patients with events. Since our measure summarizes the information on λ_E, the number of patients with $T_E < \tau$ ($\tau = 5$ in simulations) is the most important parameter. On the other hand, the patients who die from other causes serve as a nuisance parameter. We illustrate this by two scenarios which have approximately equal number of patients who die of cancer (approx 40 %), but a very different other cause hazard: we change the age distribution of the patients and the size of the baseline hazard to get 52 % and 77 % of deaths due to excess hazard, respectively. The upper limit for age never exceeds 85 and thus ensures that the probability of the patients to be still alive at the end of follow-up time (5 years) is high enough to make our interest in λ_E still sensible.

Based on the experience in the literature, we add a third scenario, which has proven as problematic with other studies in relative survival (e.g. [25] (Supp material) has shown both a huge variance and biased regression coefficient estimates in such scenarios): we consider a very low proportion of deaths due to excess hazard (36 %). Coupled with a low overall number of events (22 %), this is also an example where the sample size 500 contains very little information on λ_E (only 8% of patients die due to λ_E).
We also changed other parameters, for example the size of covariate effect β in the excess hazard. However, while these scenarios were important with other issues of relative survival (see e.g. [7]), they proved irrelevant in our case.

Simulation details Times to death from cancer (T_E) are simulated with $\lambda_E(t|X) = \lambda_{E0} \exp(\beta X)$. The basic scenario uses a fixed λ_{E0} and $\beta = 0$ for

all covariates X. Approximately 10% of patients are censored by the end of follow-up time $\tau = 5$ years. Each simulation scenario consists of 5000 repetitions, all results (bias and coverage in different scenarios) are assessed at 5 years. In all the simulations the probability that a patient is a female equals 0.5. Year of diagnosis is generated uniformly between 1st January 1990 and 31st December 1999. Age is generated uniformly from an interval whose boundaries were changed to achieve the desired proportion of deaths and deaths from other causes. To control both proportions also the baseline hazard is changed through simulations.

We first give detailed description of the simulation scenarios with no effect of covariates (Tables 3.1 and 3.3).

- Scenario where 0.22 of all the patients have events and 0.36 of these events are due to excess hazard. The baseline hazard is $5 \cdot 10^{-5}$ and age is generated uniformly from 50 to 80 years. Censoring times are generated independently, uniformly from 0 to 40 years.
- Scenario where 0.51 of all the patients have events and 0.77 of these events are due to excess hazard. The baseline hazard is $3 \cdot 10^{-4}$ and age is generated uniformly from 50 to 80 years. Censoring times are generated independently, uniformly from 0 to 40 years.
- Scenario where 0.77 of all the patients have events and 0.52 of these events are due to excess hazard. The baseline hazard is $4 \cdot 10^{-4}$ and age is generated uniformly from 80 to 85 years. Censoring times are generated independently, uniformly from 0 to 27 years.

We further present some simulation scenarios where covariate age has effect (Tables 3.2 and 3.4). We have used samples of size $n = 500$. Baseline hazard was set to $3 \cdot 10^{-4}$, age was generated uniformly from 50 to 80 years and censoring times were generated uniformly from 0 to 40 years. The proportion of events in these scenarios is around 51% and proportion of events due to excess hazard among all events is around 77%. We created four scenarios with different sizes of age effect (per year):

- small effect: $\beta_{age} = 0.007$
- medium effect: $\beta_{age} = 0.012$
- large effect: $\beta_{age} = 0.06$

- enormous effect: $\beta_{age} = 0.1$

The coefficients for the first three scenarios were obtained from Slovene cancer registry data (see [28] for details) and we made up the forth one as an example of extremely large effect.

Simulation results Results describing the bias are presented in Table 3.1. We can see that the magnitude of bias of the pseudo based approach is of the order of 10^{-4} (the standard error of the estimate is of the order of 10^{-2} in this case), the bias is therefore practically negligible. Furthermore, we note that pseudo based approach and the PP approach are almost numerically equal in all scenarios considered, the differences are negligible even in the rather extreme scenario (22% / 36 %), where the weights of the PP estimator are known to be very large. Due to this equality no further comparisons with other methods need to be done, since in terms of bias, difference between the existing estimators and variability the results will be the same as in the papers comparing the PP estimator with other approaches to estimation of net survival [7, 30, 18, 34, 33].

Table 3.1: Estimated bias of different estimators ($\tau = 5$ years); no covariate effect; different sample sizes.

		Complete data			Right censored data	
$\frac{d}{n}\left(\frac{d_E}{d}\right)$	Sample size	P	PP	h	P	PP
0.22 (0.36)	$n = 500$	$8.2 \cdot 10^{-4}$	$5.73 \cdot 10^{-4}$	$-5.37 \cdot 10^{-5}$	$9.19 \cdot 10^{-4}$	$7.18 \cdot 10^{-4}$
	$n = 2500$	$3.97 \cdot 10^{-4}$	$3.44 \cdot 10^{-4}$	$-2.33 \cdot 10^{-6}$	$3.61 \cdot 10^{-4}$	$3.26 \cdot 10^{-4}$
	$n = 5000$	$3.75 \cdot 10^{-4}$	$3.46 \cdot 10^{-4}$	$-2.44 \cdot 10^{-5}$	$4.08 \cdot 10^{-4}$	$3.87 \cdot 10^{-4}$
0.51 (0.77)	$n = 500$	$1.62 \cdot 10^{-4}$	$2.79 \cdot 10^{-4}$	$-1.90 \cdot 10^{-4}$	$1.37 \cdot 10^{-4}$	$2.24 \cdot 10^{-4}$
	$n = 2500$	$5.87 \cdot 10^{-4}$	$6.09 \cdot 10^{-4}$	$2.60 \cdot 10^{-4}$	$5.56 \cdot 10^{-4}$	$5.75 \cdot 10^{-4}$
	$n = 5000$	$6.8 \cdot 10^{-5}$	$7.75 \cdot 10^{-5}$	$-1.92 \cdot 10^{-4}$	$7.54 \cdot 10^{-5}$	$7.83 \cdot 10^{-5}$
0.77 (0.52)	$n = 500$	$1.15 \cdot 10^{-3}$	$1.67 \cdot 10^{-3}$	$-2.00 \cdot 10^{-5}$	$1.43 \cdot 10^{-3}$	$1.94 \cdot 10^{-3}$
	$n = 2500$	$7.07 \cdot 10^{-4}$	$8.02 \cdot 10^{-4}$	$-1.25 \cdot 10^{-4}$	$7.53 \cdot 10^{-4}$	$8.41 \cdot 10^{-4}$
	$n = 5000$	$1.21 \cdot 10^{-3}$	$1.25 \cdot 10^{-3}$	$-6.12 \cdot 10^{-5}$	$1.21 \cdot 10^{-3}$	$1.26 \cdot 10^{-3}$

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; P = the pseudo based approach; PP = PP estimator; h = Kaplan-Meier estimator in the hypothetical world.

Similar conclusions can be drawn from the Table 3.2. It presents results from scenarios with different sizes of age effect.

In order to compare the two formulae for estimation of variance we consider the coverage of the confidence intervals. Since the pseudo based approach is practically unbiased, the coverage entirely reflects the precision of the two estimators of variance. The simulation results for three different scenarios

Table 3.2: Estimated bias of different estimators ($\tau = 5$ years); different covariate effects; $n = 500$.

	Complete data			Right censored data	
	P	PP	h	P	PP
small ef.	$1.71 \cdot 10^{-5}$	$1.31 \cdot 10^{-4}$	$1.10 \cdot 10^{-5}$	$7.05 \cdot 10^{-5}$	$1.61 \cdot 10^{-4}$
medium ef.	$3.24 \cdot 10^{-4}$	$4.37 \cdot 10^{-4}$	$-5.33 \cdot 10^{-5}$	$2.80 \cdot 10^{-4}$	$3.88 \cdot 10^{-4}$
large ef.	$-5.03 \cdot 10^{-4}$	$-3.81 \cdot 10^{-4}$	$-3.96 \cdot 10^{-4}$	$-5.02 \cdot 10^{-4}$	$-4.07 \cdot 10^{-4}$
enormous ef.	$-3.97 \cdot 10^{-4}$	$-2.89 \cdot 10^{-4}$	$-3.51 \cdot 10^{-4}$	$-4.39 \cdot 10^{-4}$	$-3.45 \cdot 10^{-4}$

P = the pseudo based approach; PP = PP estimator; h = Kaplan-Meier estimator in the hypothetical world; proportion of events is 0.51; of these 0.77 are due to excess hazard.

with different proportions of deaths and different proportions of deaths due to excess hazard among all deaths for three different sample sizes are given in Table 3.3. They indicate that the precise formula gives coverages that are close to the nominal value of 0.95 in all three scenarios with the most extreme scenario (the smallest proportion of events) tending to have a slightly too low coverage. This is in line with the previous work, which indicates disproportionately large variability with small proportions of excess hazard events (e.g. Figure 1 in the Supplementary material of [25]), in our case, this large variability seems to be underestimated. The approximate formula gives coverages that are constantly below the nominal level (see Table 3.3) but the discrepancies are minimal in the two realistic scenarios with the higher proportion of deaths. As a standard to compare to, we also report the coverage of the PP estimator, which is close to the nominal value in all cases.

Comparing all the results in Table 3.3, it is clear that the problems do not lie in small samples, instead, the proportion of excess hazard deaths among all deaths is crucial for the properties of our estimators. Table 3.4 presents results from scenarios with different sizes of age effect. The results are in line with those from Table 3.3, an exception being the last line where both the precise formula and the PP estimator give too wide confidence intervals. We can notice that in this case also the Kaplan-Meier in hypothetical world has too large coverage and we cannot expect the other two methods that operate on less informative data to outperform it. Furthermore, the trend seems to be that the coverage increases (and this does not bother us so much as coverage that would be too low). Also, this scenario is less likely to occur in practice due to extremely large effect. All in all, we can conclude that this is not a problem for the practical use.

Table 3.3: Estimated coverages of CI ($\tau = 5$ years, nominal value 0.95); no covariate effect; different sample sizes.

$\frac{d}{n}\left(\frac{d_E}{d}\right)$	0.22 (0.36)			0.51 (0.77)			0.77 (0.52)		
estimator	*P(pr)*	*P(apr)*	*PP*	*P(pr)*	*P(apr)*	*PP*	*P(pr)*	*P(apr)*	*PP*
$n = 500$	0.9404	0.9108	0.947	0.954	0.9468	0.9546	0.9454	0.9424	0.9458
$n = 2500$	0.9378	0.9128	0.949	0.9442	0.9382	0.9464	0.956	0.9544	0.9562
$n = 5000$	0.9386	0.9116	0.9488	0.943	0.9366	0.9462	0.9482	0.946	0.9498

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *P (pr)* = conf. intervals for the pseudo based approach, variance estimated by the precise formula; *P (apr)* = conf. intervals for the pseudo based approach, variance estimated by the approximate formula; *PP* = conf. intervals for PP estimator.

Table 3.4: Estimated coverage of confidence intervals ($\tau = 5$ years; nominal value 0.95); different covariate effects; $n = 500$.

	Complete data				Right censored data		
	P (pr)	*P (apr)*	*PP*	*h*	*P (pr)*	*P (apr)*	*PP*
small ef.	0.9486	0.9422	0.9504	0.95	0.9486	0.9408	0.9506
medium ef.	0.9442	0.9364	0.9476	0.9542	0.9432	0.9354	0.946
large ef.	0.956	0.9386	0.9598	0.961	0.9562	0.9394	0.9624
enormous ef.	0.9696	0.9466	0.9766	0.9738	0.97	0.9456	0.9758

P (pr) = conf. intervals for the pseudo based approach, variance estimated by the precise formula; *P (apr)* = conf. intervals for the pseudo based approach, variance estimated by the approximate formula; *PP* = conf. intervals for PP estimator; h = conf. intervals for Kaplan-Meier estimator in the hypothetical world; proportion of events is 0.51; of these 0.77 are due to excess hazard.

3.3.4 Example

Computers are useless. They can only give you answers.

Pablo Picasso

We now explore the properties of the estimator on real data of Slovene female colon cancer patients (see Subsection 1.4.2 for details). Figure 3.3 (left side) presents their 20 year survival curve. This sample was chosen on purpose - as an example of a truly bizarre behaviour we can get with routine use of the estimator, a point raised by [9] and [18]. To understand the reasons for this behaviour, note first, that despite the weird behaviour both the PP estimator and our new proposal actually give the same estimates. The same would be true also if we had no censoring, i.e. in the case, where the pseudo based approach follows directly from the definition of the measure (see Subsection 3.3.2 for details). It is thus clear that this behaviour is not a property of the PP estimator or the pseudo based approach, but rather the property of the measure. The weights that cause the variable behaviour of the estimators are present already in the definition of the measure and hence affect any

estimator of this measure. The oldest individuals in our sample were aged 85 and their probability of still being alive after 20 years ($S_P(20)$) is of order 10^{-6}, we hence cannot expect to have any information on their $\widetilde{\lambda}_E$ in the data. Before dropping from this sample, these individuals can have very large weights - in the case of our sample, there were 267 patients still at risk at 17 years, with two patients having weights of size 50 (the median weight at that time is 1.3), these two patients almost entirely control the behaviour of the estimator at that time.
Some authors propose alternative estimators to deal with such data (age-stratified Ederer II, model based estimator, etc., see e.g. [9]), i.e. introducing assumptions to deal with the problem of little or no information available in the data. As the above example clearly illustrates, the large variability is not a problem of a specific estimator but rather a problem of the measure. Instead of reasoning why the assumptions made (that are almost always untestable) are acceptable, we focus on understanding this property of the measure and the limitations it brings with.

It is not clear that intelligence has any long-term survival value.

Stephen Hawking

The measure is a function of $\widetilde{\lambda}_E$ of all patients. With the probability of still being alive in the population $\widetilde{S}_P$ getting very low, we cannot expect to get any information on $\widetilde{\lambda}_E$ from the data. One way to move forward is to set some assumptions, but, apart from being untestable, any assumptions about the value of $\widetilde{\lambda}_E$ for very old patients are also not at all of practical interest. We therefore propose one should limit to a cohort for which the value of $\widetilde{\lambda}_E$ is of interest throughout the follow-up interval. In our example - if the interest lies in 20 year survival, it seems reasonable to limit ourselves to patients under 70 at diagnosis, thus ensuring that their probability of still being alive in the population after 20 years is large enough to allow estimation. The graph of the resulting 815 women is presented on the right side of Figure 3.3. Again, we see that the values of the PP and the estimator $\widehat{\widetilde{S}}_E(t)$ are practically identical, the same is true for the confidence intervals. When limiting the analysis to follow-up for which the measure is meaningful, both estimators (PP and the pseudo based approach) shall have reasonable properties in terms of variance.

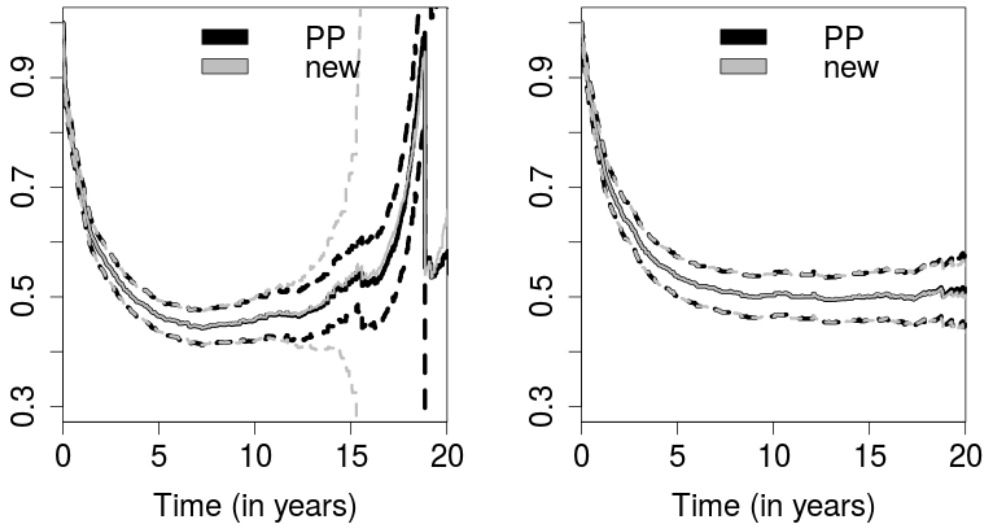

Figure 3.3: The PP and the pseudo based approach on the data of Slovene female colon cancer patients; left: complete data; right: patients under 70.

3.4 Discretely measured time

In the previous sections we worked with the assumption that time is measured continuously, i.e. for each patient his exact event or censoring time is known.

The data are sometimes grouped so that we know only in which time interval (e.g. week, month) a patient had an event or was censored and not the precise day when this happened. The time points at which we have data are common for all the patients. This can happen either because of poor data collection quality or due to increasing awareness about personal data protection (more likely in the modern world and in the future). Such data presents new challenges for data analysts. Techniques developed for continuously measured time may perform poorly when used on discretely measured time. The PP estimator is no exception, it gives biased estimates when used with discretely measured time. This has already been noticed in [34].

The purpose of this section is to extend the pseudo based approach to data with discretely measured time. We start from the estimator (3.1), but we have to choose appropriate estimators $\widehat{S}_{Oi}(t)$. As before, we can use pseudo observations but we have to replace Kaplan-Meier estimator with another estimator that has nicer properties when used on discretely measured time.

3.4.1 Life table estimator

Let $t_0 = 0 < t_1 < \cdots < t_k < t_{k+1} = \infty$ be a division of the time axis into $k+1$ intervals and let n_j denote the number of patients still at risk at the beginning of the interval $[t_{j-1}, t_j)$. Let d_j and c_j denote the number of events and censored observations in the interval $[t_{j-1}, t_j)$ and let $n'_j = n_j - c_j/2$ denote the modified number of patients at risk. Let us also define $\frac{d_j}{n'_j} = 1$ whenever $n_j = 0$. The standard life table estimator is defined as [20]

$$\widehat{S}_{LT}(t_j) = \prod_{i \le j} \Big(1 - \frac{d_i}{n'_i}\Big). \tag{3.17}$$

Its variance is estimated by

$$\widehat{\mathrm{var}}\widehat{S}_{LT}(t_j) = \widehat{S}_{LT}(t_j)^2 \sum_{i \le j} \frac{d_i}{n'_i(n'_i - d_i)}. \tag{3.18}$$

Formulas (3.17) and (3.18) are very similar to those for Kaplan-Meier estimator and the estimator of its variance, the only difference being the modified number of patients at risk. This modification is achieved by subtracting one half of the number of censored patients to take into account that those that are censored could die before the end of the interval. This is an ad hoc correction which corresponds to uniform censoring mechanism. Other censoring mechanisms would correspond to different modifications of the number of patients at risk. Since the censoring mechanism is typically not known, the common choice is to substract one half of the censored patients to get the effective number of patients at risk. This modification usually works reasonably well in practice, in particular if the width of the grouping intervals $t_{i+1} - t_i$ is not too large.

3.4.2 Estimation of marginal relative survival and estimation of the variance of its estimator

We propose to use the standard life table estimator $\widehat{S}_{LT}$ to calculate pseudo observations. We plug them into the formula (3.1) to get an estimator for marginal relative survival. Formula (3.18) is obtained by replacing the number of patients at risk n_i in the estimator of variance of Kaplan-Meier estimator with the effective number of patients at risk n'_i. We propose the same ad

hoc correction of formulas (3.14) and (3.15) to obtain estimators of variance for the discretely measured time.

3.4.3 Simulation study

We use simulations to explore the behaviour of estimator (3.1) on discretely measured time. We change four parameters across different simulation scenarios: the width of the aggregation window $t_{i+1} - t_i$, the proportion of censored observations, censoring distribution and the effect size. When the aggregation window is smaller, less information is lost and we expect the bias of the PP estimator to be smaller. When there are no censored observations, the ad hoc correction term from the estimator (3.17) vanishes and there is no additional assumption about the censoring mechanism. However, when the proportion of censoring increases, this assumption can gain importance and its violation can have larger impact on the results. The ad hoc correction used to obtain the effective number of patients at risk introduces the assumption that censoring is uniform and violations of this assumption could affect the properties of estimator.

Simulation details

Times to death from cancer are simulated from the exponential model for hazard $\lambda_E(t|X) = \lambda_{E0} \exp(\beta X)$ with $\lambda_{E0} = 3 \cdot 10^{-4}$. Covariate X is age (uniform between 50 and 80 years) and β is either 0 (no effect) or 0.1 (enormous effect). Time of diagnosis is generated uniformly between the beginning of 1990 and the end of 1999. A patient is a female with probability 0.5 and τ is 5 years. Observed times $\widetilde{T}_i$ are then aggregated using three different values for the width of the aggregation window: 1 year, 3 months and 1 month. Censoring distribution is either uniform or exponential. Proportion of censored observations is 9%, 34% to 36% or 54% to 60%. When there is no (an enormous) effect of age, the proportion of all events equals 48%, 40%, 32% (49%, 41%, 34%) and the order is the same as with the proportions of censored observations. Of these events, 77% (82%) can be contributed to cancer. Sample size is 1000 and 1000 simulation runs are performed in each case.

Simulation results

We will focus on estimated bias and coverage of confidence intervals. There are other measures (such as mean square error) that reflect the performance of an estimator (and can be used to compare (biased) estimators) but our primary goal is to have smaller bias and correct coverage.

Tables 3.5 and 3.6 present the estimated bias of both the PP estimator and the pseudo based approach for two different censoring distributions. The pseudo based approach gives estimates that are on average much less biased (roughly speaking ten times smaller) compared to the PP estimator. Important differences can be observed with higher proportions of censoring and wider aggregation windows where bias of the PP estimator can be much larger than 1% and the bias of pseudo based approach remains under 1%. Censoring distribution does not seem to influence the bias.

Table 3.5: Estimated bias of $\widehat{S}(\tau)$ ($\times 10^{-2}$; discretely measured time, uniform censoring, $\tau = 5$ years, $n = 1000$).

		$t_{i+1} - t_i = 1$ year		$t_{i+1} - t_i = \frac{1}{4}$ year		$t_{i+1} - t_i = \frac{1}{12}$ year	
β_{age}	estimator	PP	P_{LT}	PP	P_{LT}	PP	P_{LT}
0	cens. level: 9%	1.707	0.094	0.428	0.013	0.248	0.115
	cens. level: 36%	4.030	0.232	0.901	-0.083	0.425	0.072
	cens. level: 60%	7.615	0.271	1.981	-0.077	0.691	0.006
0.1	cens. level: 9%	1.571	0.061	0.360	0.005	0.115	0.001
	cens. level: 35%	3.597	-0.016	0.872	-0.023	0.363	0.064
	cens. level: 57%	7.053	0.181	1.963	0.100	0.563	-0.097

$t_{i+1} - t_i$ = width of time window; β_{age} = effect of age per year; PP = bias of the PP estimator; P_{LT} = bias of the pseudo based approach (which is calculated using the life table estimator).

Table 3.6: Estimated bias of $\widehat{S}(\tau)$ ($\times 10^{-2}$; discretely measured time, exponential censoring, $\tau = 5$ years, $n = 1000$).

		$t_{i+1} - t_i = 1$ year		$t_{i+1} - t_i = \frac{1}{4}$ year		$t_{i+1} - t_i = \frac{1}{12}$ year	
β_{age}	estimator	PP	P_{LT}	PP	P_{LT}	PP	P_{LT}
0	cens. level: 9%	1.675	0.034	0.335	-0.062	0.054	-0.092
	cens. level: 34%	3.834	0.224	0.995	0.106	0.276	-0.028
	cens. level: 56%	6.563	0.673	1.402	-0.165	0.632	0.105
0.1	cens. level: 9%	1.579	0.126	0.380	0.018	0.137	0.020
	cens. level: 35%	3.634	0.167	0.987	0.115	0.296	0.003
	cens. level: 54%	6.293	0.528	1.524	0.028	0.575	0.075

$t_{i+1} - t_i$ = width of time window; β_{age} = effect of age per year; PP = bias of the PP estimator; P_{LT} = bias of the pseudo based approach (which is calculated using the life table estimator).

Tables 3.7 and 3.8 present the estimated coverages of the confidence intervals

(for two different censoring distributions). We realize that the coverage lower than the nominal level can arise from the bias of the estimator.

The approximative formula for the variance gives lower coverages when compared to the precise formula for the variance of the pseudo based approach - this is in line with the results on the continuously measured time where the approximative formula also slightly underestimates the variance. Note that with discretely measured time there are fewer time points and therefore the precise formula is less computationally demanding - there is not such a big need for the approximative formula in this case.

When the aggregation window is one month, both the PP estimator and the precise formula for the variance of the pseudo based approach give similar coverages that are close to the nominal level. However, when the aggregation window is wider and the proportion of censoring increases, the coverages of the PP estimator get below the nominal level. Precise formula for the variance of the pseudo based approach still gives coverages that are close to the nominal level.

Again, the censoring distribution has no important impact on the coverage of confidence intervals obtained by the precise formula for the variance of the pseudo based estimator: when the censoring mechanism is uniform the coverages are a bit closer to the nominal level but the differences are minimal and can be observed only with yearly grouping of the data. Interestingly, with wider aggregation windows and the highest proportion of censoring the coverages of confidence intervals obtained by the PP estimator are higher for the exponential censoring distribution. Coverage is too high with all three estimation approaches when the effect size is enormous ($\beta_{age} = 0.1$) and the proportion of censoring is only 9%.

Bias of the pseudo based approach is smaller or equal compared to the PP estimator. Furthermore, the results obtained by the precise formula for the variance of the pseudo based approach are closest to the nominal value and therefore this seems to be the most reliable option. This formula also gets less computationally intensive since there are less different time points with discretely measured data. This pair of estimators thus outperforms the PP estimator on discretely measured data.

Table 3.7: Estimated coverage of confidence intervals (discretely measured time, uniform censoring, $\tau = 5$ years, $\alpha = 0.05$, $n = 1000$).

		$t_{i+1} - t_i = 1$ year			$t_{i+1} - t_i = \frac{1}{4}$ year			$t_{i+1} - t_i = \frac{1}{12}$ year		
β_{age}	estimator	PP	$P_{LT,p}$	$P_{LT,a}$	PP	$P_{LT,p}$	$P_{LT,a}$	PP	$P_{LT,p}$	$P_{LT,a}$
	cens. level: 9%	0.849	0.955	0.951	0.942	0.942	0.938	0.942	0.944	0.939
0	cens. level: 36%	0.536	0.948	0.937	0.926	0.935	0.927	0.949	0.948	0.944
	cens. level: 60%	0.154	0.949	0.943	0.876	0.951	0.946	0.932	0.933	0.917
	cens. level: 9%	0.904	0.975	0.962	0.970	0.967	0.950	0.976	0.966	0.947
0.1	cens. level: 35%	0.614	0.959	0.938	0.948	0.953	0.928	0.965	0.957	0.930
	cens. level: 57%	0.153	0.952	0.932	0.895	0.949	0.927	0.955	0.953	0.919

$t_{i+1} - t_i$ = width of time window; β_{age} = effect of age per year; PP = coverage of the CI using the PP estimator; $P_{LT,p}$ = coverage of the CI using the pseudo based approach (which is calculated using the life table estimator), variance is estimated by the precise formula; $P_{LT,a}$ = coverage of the CI using the pseudo based approach (which is calculated using the life table estimator), variance is estimated by the approximate formula.

Table 3.8: Estimated coverage of confidence intervals (discretely measured time, exponential censoring, $\tau = 5$ years, $\alpha = 0.05$, $n = 1000$).

		$t_{i+1} - t_i = 1$ year			$t_{i+1} - t_i = \frac{1}{4}$ year			$t_{i+1} - t_i = \frac{1}{12}$ year		
β_{age}	estimator	PP	$P_{LT,p}$	$P_{LT,a}$	PP	$P_{LT,p}$	$P_{LT,a}$	PP	$P_{LT,p}$	$P_{LT,a}$
	cens. level: 9%	0.865	0.963	0.959	0.939	0.945	0.938	0.959	0.955	0.950
0	cens. level: 34%	0.565	0.939	0.933	0.921	0.955	0.951	0.953	0.958	0.951
	cens. level: 56%	0.300	0.925	0.920	0.934	0.940	0.934	0.923	0.927	0.915
	cens. level: 9%	0.891	0.974	0.964	0.978	0.976	0.965	0.980	0.978	0.961
0.1	cens. level: 35%	0.592	0.961	0.944	0.971	0.977	0.960	0.969	0.958	0.944
	cens. level: 54%	0.296	0.943	0.920	0.930	0.954	0.934	0.957	0.948	0.924

$t_{i+1} - t_i$ = width of time window; β_{age} = effect of age per year; PP = coverage of the CI using the PP estimator; $P_{LT,p}$ = coverage of the CI using the pseudo based approach (which is calculated using the life table estimator), variance is estimated by the precise formula; $P_{LT,a}$ = coverage of the CI using the pseudo based approach (which is calculated using the life table estimator), variance is estimated by the approximate formula.

Chapter 4

Comparison of net survival curves

Comparison is the death of joy.

Mark Twain

In this chapter we focus on comparison of marginal relative survival and net survival between different groups. Even though the title of this chapter includes only the latter measure, we write about both of them, since they start from a similar model but with a different amount of assumptions. The estimates of the two measures are practically identical (see Section 3.3) but have different interpretation.

4.1 Methods for comparison

In 2016 Grafféo et al. [13] introduced a log-rank type test that combines the ideas of the log-rank test and the weighting of the PP estimator to adjust the counting process N and the at risk process Y. The paper nicely formulates the test statistic, develops the required theoretical results and uses simulations to check them. However, comparison with other methods is missing and is thus unclear what is the right place for this test among existing methods.

The log-rank type test was designed to compare net survival curves, i.e. to test the null hypothesis $\lambda_{E,1}(t) = \cdots = \lambda_{E,k}(t)$, where $\lambda_{E,i}(t) = \lambda_E(t|X = i)$

and X is a covariate which splits the data into k groups. The ideal data to test this null hypothesis would be the pairs $(\widetilde{T}_{Ei}, \delta_{Ei})$, i.e. the *hypothetical world* data. These data cannot be available in real world, where there is no way to exclude other causes, but we can use it in theory and simulations to better understand the properties of the log-rank type test and to compare them to the properties of the classical log-rank test. The real world data present the competing risks setting from which we wish to extract the information about $\tilde{\lambda}_E$, whereas the hypothetical world data present the simpler framework where patients are subject only to one hazard and thus the basic survival methodology (Kaplan-Meier estimator, Cox model) can be directly used.

We shall thus first look at the relations in the hypothetical world and then try to find similar relations in the real world.

Hypothetical world data

Using the counting process notation, we let $N_{E,i}(t)$ denote the counting process for individual i: $N_{E,i}(t) = I(\widetilde{T}_{E,i} \leq t, T_{E,i} \leq C_i)$ and $Y_{E,i}(t)$ denote the at risk process for each individual: $Y_{E,i}(t) = I(\widetilde{T}_{E,i} \geq t)$. Further, we use $N_{E,h}(t)$ and $Y_{E,h}(t)$ for the sum of $N_{E,i}$ and $Y_{E,i}$ for all individuals i belonging to each of the subgroups $h = 1, \ldots, k$. The test statistic compares the observed and the expected number of events in group h at each time point. The observed number of events at each time u is denoted as $dN_{E,h}(u)$, the expected number of events is calculated from the total number of deaths at that time($dN_{E,\cdot}(u)$) as the proportion corresponding to the ratio between the number of individuals at risk in group h ($Y_{E,h}(u)$) and the total number of individuals still at risk ($Y_{E,\cdot}(u)$):

$$Z_h(\tau) = \int_0^\tau I\big(Y_{E,\cdot}(u) > 0\big)\left[dN_{E,h}(u) - \frac{Y_{E,h}(u)}{Y_{E,\cdot}(u)} dN_{E,\cdot}(u)\right], \tag{4.1}$$

where τ is the follow-up time. The test statistic is then calculated as

$$U = Z^T \widehat{\Sigma}^{-1} Z,$$

where $Z = (Z_1(t), \ldots, Z_{k-1}(t))'$ and $\widehat{\Sigma}$ is the estimated variance-covariance matrix, see [11] for details. Under the null hypothesis, the test statistic can be shown to be asymptotically χ^2_{k-1} distributed.
The derivation of the log-rank test requires the hazard $\lambda_{E,h,i}$ to be equal for all individuals i in a certain group h, we shall refer to this property as the homogeneity of the hazards. If this is not true and there is another

categorical variable S which explains the differences within each subgroup, one can use the stratified log-rank test, where the homogeneity property is required to hold only within the strata of each group. We calculate the Z_h value in each stratum separately

$$Z_{h,s}(\tau) = \int_0^\tau I\big(Y_{E,\cdot,s}(u) > 0\big)\left[dN_{E,h,s}(u) - \frac{Y_{E,h,s}(u)}{Y_{E,\cdot,s}(u)} dN_{E,\cdot,s}(u)\right] \tag{4.2}$$

and then sum over all m strata to get the test statistic (see [11] for more details)

$$U = \big(\sum_{s=1}^{m} Z_s\big)^T \big(\sum_{s=1}^{m} \widehat{\Sigma}_s\big)^{-1} \big(\sum_{s=1}^{m} Z_s\big),$$

where $Z_s = (Z_{1,s}(t), \ldots, Z_{k-1,s}(t))$.

Real world data

With the real world data, we wish to test the same null hypothesis of equal excess terms, but the number of cause-specific deaths ($dN_{E,h}$ in formula (4.1)) and the number at risk in the hypothetical world ($Y_{E,h}$ in formula (4.1)) cannot be directly observed. They thus have to be estimated by the help of population tables, we shall denote these estimates by $\widehat{dN}_{E,h}$ and $\widehat{Y}_{E,h}$ respectively. Let (N_i, Y_i) denote the equivalents of the processes $(N_{E,i}, Y_{E,i})$ defined from the observed data $(\widetilde{T}_i, \delta_i)$ (see Subsection 3.2.2 for details). We use these data and merge it with the population mortality data to estimate the number of deaths from cancer (see [26] for details):

$$\widehat{dN}_{E,i}(t) = \frac{dN_i(t)}{S_{P,i}(t)} - \int_0^t \widehat{Y}_{E,i}(u)\lambda_{P,i}(u)du, \tag{4.3}$$

where $\widehat{Y}_{E,i}(t) = \frac{Y_i(t)}{S_{P,i}(t)}$. In this estimation, we follow the idea of the PP estimator [26]: the total number of events must be diminished by the number of expected deaths in the population (second term on the right of (4.3)) and both the observed number of deaths and number at risk must be weighted (by $S_{P,i}$) to properly represent the numbers we would observe in the hypothetical world where no one dies from other causes. However, this weighting properly describes the quantities from the hypothetical world only when the assumptions of the PP estimator are fulfilled. This is often not true (because the assumption that population mortality tables reflect the other cause mortality and the assumption (2.10) are not valid). Therefore, the null hypothesis of

the test is $\widetilde{\lambda}_{E,1}(t) = \cdots = \widetilde{\lambda}_{E,k}(t)$, i.e. the test compares marginal relative survival between groups. Note that $\widetilde{\lambda}_{E,i} = \lambda_{E,i}$ under conditions presented in Section 2.4.

The test statistic is calculated using [13]:

$$Z_h(\tau) = \int_0^\tau I\big(\widehat{Y}_{E,\cdot}(u) > 0\big)\left[\widehat{dN}_{E,h}(u) - \frac{\widehat{Y}_{E,h}(u)}{\widehat{Y}_{E,\cdot}(u)}\widehat{dN}_{E,\cdot}(u)\right]. \tag{4.4}$$

Similarly, the stratified version is calculated as

$$Z_{h,s}(\tau) = \int_0^\tau I\big(\widehat{Y}_{E,\cdot,s}(u) > 0\big)\left[\widehat{dN}_{E,h,s}(u) - \frac{\widehat{Y}_{E,h,s}(u)}{\widehat{Y}_{E,\cdot,s}(u)}\widehat{dN}_{E,\cdot,s}(u)\right]. \tag{4.5}$$

4.1.1 Properties and interpretation of the log-rank type test

The log-rank test on hypothetical world data and the log-rank type test on real world data have the same null hypothesis (when the assumptions of the PP estimator are valid for the real world data) and differ only due to the different data available. We can therefore expect similar behaviour, but a smaller power in the case of real world data, where some of the information cannot be observed. Since the interest lies in the comparison of the excess term $\widetilde{\lambda}_E$, we can expect the proportion of events due to excess term to crucially affect the power.

Log-rank test and regression models

To further understand the properties and the interpretation of the log-rank type test we compare it to its main alternative - a regression model. The standard log-rank test statistics comparing two groups defined by the binary covariate X is identical to the score test statistic in a univariate Cox model containing covariate X (when no ties are present) [2]. Therefore, the two approaches have the same properties and can be interpreted in the same way. This means that with hypothetical world data, the score test statistic of the null hypothesis $H_0 : \beta_X = 0$ in the Cox model

$$\lambda_E(t, x) = \lambda_{E_0}(t)e^{\beta_X x} \tag{4.6}$$

is identical to the log-rank test statistic.
The analogous model with the real world data $(\widetilde{T}_i, \delta_i)$ is the additive model

$$\lambda_O(t,x) = \widetilde{\lambda}_P(t,x) + \widetilde{\lambda}_{E_0}(t)e^{\beta_X x} \tag{4.7}$$

Again, the null hypothesis $H_0 : \beta_X = 0$ is of interest and this null hypothesis is clearly equivalent to the null hypothesis of the log-rank type test, i.e. $\widetilde{\lambda}_{E,1}(t) = \widetilde{\lambda}_{E,2}(t)$, where the two groups are defined by the binary covariate X.
In order to expect similar behaviour of the additive model and the log-rank type test when testing this hypothesis, the assumptions required by both methods should also be the same. This implies that the $\widetilde{\lambda}_{E_0}(t)$ in (4.7) should be left completely unspecified, since the log-rank type test also makes no distributional assumption within each group. Unfortunately, the idea of using partial likelihood in the framework of the relative survival does not work, since the baseline hazard does not cancel out (due to $\widetilde{\lambda}_P$ in (4.7)). Therefore, the $\widetilde{\lambda}_{E_0}(t)$ in model (4.7) is usually defined as a function of a few parameters - it has been initially defined with a stepwise constant hazard function [10] and many alternative parametric specifications have been proposed since [12, 19, 21]. The fully parametric model specified in this way can be fitted using maximum likelihood. On the other hand, only two semi-parametric approaches allowing $\widetilde{\lambda}_{E_0}(t)$ to be left unspecified have been proposed [25, 32]. Both approaches require some smoothing when fitting (hence making some weak assumptions about the baseline hazard form), therefore, these models cannot be expected to give identical results as the log-rank type test either. It can be quickly seen that the test statistics and the p values are not identical, therefore, the question is to what extent the log-rank type test and the regression models behave in the same way and can be interpreted in the same way.

The stratified vs the non-stratified version

The derivation of the distribution of the log-rank test statistic under the null hypothesis requires $\lambda_E(t)$ to be equal for individuals within a certain subgroup defined by X [13]. This may not be true in practice, hence a stratified version is proposed to take this inhomogeneity into account. The stratified log-rank test implies that the groups formed by X are not compared on the whole sample but rather in strata defined by a covariate S. Then, the results from all strata are pulled together to form a single test statistic value. The direct analog of this approach with the hypothetical world data

is the stratified Cox model

$$\lambda_E(t, x, s) = \lambda_{E_0,s}(t)e^{\beta_X x}$$

in which the baseline hazard is allowed to differ between strata, but a common coefficient β_X describing the effect of X is estimated. In the hypothetical world, where the framework of the classical survival analysis is used, the stratified log-rank test and the score test of the stratified regression model give identical results. On the other hand, the estimation of β_X in a Cox model stratified by S can be compared to a multivariate model containing both X and S. If the covariate S satisfies the proportional hazards assumption (i.e. $\lambda_{E_0,s}(t) = \lambda_{E_0}(t)\exp(\beta_S s)$), β_X has the same interpretation with both the stratified and the multivariate model. Note that if X and S are not independent, the common analogy of the stratified and the multivariate regression model implies that the interpretation of the stratified model is importantly different from the non-stratified version (i.e. univariate model). To draw the analogy further, say we wish to compare two groups defined by X with the hypothetical world data. If hazards within these groups are not homogeneous, but depend on S, the data follow a model with both covariates (X and S) and the Cox model fit will not be perfect if S is omitted. The estimated coefficient for X will shrink toward zero, and the power for testing the null hypothesis $H_0 : \beta_X = 0$ shall be lower. Since the Cox model score test statistic and log-rank test statistic are equal, loss of power shall also occur with the log-rank test if S is not taken into account and the non-stratified version of the test is used. The same can be then expected also when using the relative survival methodology.
The stratified log-rank type test may therefore be used for two reasons: to correct for the fact that hazard is not homogeneous within subgroups (an assumption needed for theoretical derivations and to potentially increase power) and to compare subgroups conditional on a second covariate S.

4.1.2 Simulations

Based on theoretical relationships given in Subsection 4.1.1, we can formulate two main issues to be explored with simulations:

- How does the log-rank type test relate to the additive model? We know that the two tests address the same null hypothesis, but their

test statistics are not the same. The questions to be answered are: can we expect the same size under the null hypothesis, do the tests have the same power with different alternatives?

- Can the non-stratified version of the log-rank type test be used even if comparing groups with non-homogeneous hazards? How does the homogeneity assumption affect the size of the test, how is the power affected, when should the stratified test be used instead?

Simulation design

The scenarios of all simulations have some common properties, that will allow for clear comparisons:

- We fix the data set size to 1000, since small sample behaviour is not of primary interest. 5000 simulation runs are performed in each scenario.
- We censor all the individuals after 10 years of observation. We do not censor any data before that time since censoring is not a crucial issue we would like to research.

In the simulations both the log-rank type test and the additive regression model address the same question defined in the hypothetical world, any difference that may be observed between the two methods must come from the different amount of information available in the real world. We thus design the simulations in a way that allows us to carefully study these differences. All the simulations are simplified so that the cause for the properties can be tracked and different grades of effects are used in several cases to show how a certain property gains importance.

- We use three distributions for demographic variables D: sex is always balanced, calendar year is uniform between 1990 and 2000, age is a mixture of two uniform distributions , i.e. we simulate age from $U(a, a+30)$ for 70% of patients and from $U(a+10, a+30)$ for the remaining 30% of patients to get more older patients.
- To make the scenarios comparable in real world data, we set the baseline excess hazard λ_{E_0} to get a similar overall number of deaths - approximately half of the individuals die of any cause in the period of observation (10 years).

- The three different scenarios have
  - $\lambda_{E_0} = 7.6 \times 10^{-5}$, $a = 50$: this corresponds to 41% of deaths due to excess term among all deaths,
  - $\lambda_{E_0} = 1.15 \times 10^{-4}$, $a = 45$: this corresponds to 62% of deaths due to excess term among all deaths (a simplified version of example on myocardial infarction),
  - $\lambda_{E_0} = 1.75 \times 10^{-4}$, $a = 25$: this corresponds to 93% of deaths due to excess term among all deaths.

The main factor we have thus changed in the scenarios is the importance of the excess hazard λ_E compared to the population hazard λ_P. We measure this importance with the proportion of deaths due to excess hazard, i.e. the proportion of patients that died ($\delta_i = 1$) in which $T_{Ei} < T_{Pi}$.

In terms of the covariates X and S, the main differences in the real world when compared to the hypothetical world could arise from the question whether a covariate is (not) in the population tables (we denote this by $X \in D$ or $X \notin D$). We also change the number of groups defined by both covariates and, in case of binary variables, make their distribution balanced (the groups occurring with equal probability, denoted as *bal*) or imbalanced (the group with the higher value of X occuring with probability 0.9).

We compare four different test statistics.

- The log-rank type test (denoted as LRt) calculated on the real world data.
- The results of a coefficient test in the semi-parametric additive model (sAM). The EM model [25] was chosen since the functions are readily available in the `relsurv` [23] package. The Wald test is used in all cases.
- The results of a Wald test for a coefficient in a fully-parametric additive model (fAM) by (4.7). This model is modelled as in [10] with λ_{E_0} being constant. It is used as the reliable option since the semi-parametric models have not been used much in practice. We use it primarily to double-check our simulations. Its results are not directly comparable to LRt in terms of power since it works under the additional assumption that the baseline excess hazard is constant (which is true in our simulations).

- The log-rank test (denoted as LRh) calculated on the hypothetical world data. This test is used as an ideal situation to which the other tests can only approach. It is also used to double-check our simulations. The results of the Cox model coefficient test in the hypothetical world are not reported, since they are practically the same.

Technical note: Wald test is used for practical reasons - it is available in the software for both semi-parametric and fully parametric additive model. It is also asymptotically equivalent to the score test and the differences should be minimal since we are using large samples.

4.1.3 Simulation results

In this section we present the simulations results. A large scale simulation study has been performed but only the results that bring interesting insight are reported in the tables.

Log-rank type test and regression - comparison of the size

First, we compare the behaviour of the log-rank type test with that of the additive model. We start with a comparison of the size of the two tests under the null hypothesis. In Table 4.1 the results of several tests are compared. For each test, we report the proportion of simulation runs in which the null hypothesis was rejected, i.e. the p value was below 0.05.

Results: We can see (Table 4.1) that the size of the log-rank type test (LRt) is close to the nominal value and does not seem to be liberal in in any of the scenarios. The same can be claimed for the fully parametric model (fAM) and the semi-parametric model (sAM), the only exception is the scenario with only 41% of deaths due to excess hazard, where the size of the latter can be above its nominal value (a problem noted already in [25]). The log-rank test in the hypothetical world (LRh) serves as a check that the simulations are properly conducted.

Table 4.1: Comparison of the log-rank type test and the additive model: size.

	$\frac{d_E}{d} = 0.62$				$\frac{d_E}{d} = 0.93$			
	LRt	*sAM*	*fAM*	*LRh*	*LRt*	*sAM*	*fAM*	*LRh*
$X \in D$, bal, bin (sex)	0.046	0.050	0.045	0.052	0.047	0.049	0.049	0.047
$X \notin D$, bal, bin	0.049	0.052	0.046	0.053	0.047	0.048	0.048	0.049
$X \notin D$, imbal, bin	0.050	0.053	0.050	0.052	0.049	0.045	0.045	0.047
$X \notin D$, bal, 4 grps	0.052	0.052	0.041	0.050	0.046	0.046	0.046	0.047
	$\frac{d_E}{d} = 0.41$							
	LRt	*sAM*	*fAM*	*LRh*				
$X \in D$, bal, bin (sex)	0.046	0.071	0.047	0.051				
$X \notin D$, bal, bin	0.051	0.066	0.045	0.048				
$X \notin D$, imbal, bin	0.057	0.054	0.049	0.047				
$X \notin D$, bal, 4 grps	0.050	0.067	0.035	0.055				

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *LRt* = log-rank type test; *sAM* = semi-parametric additive model; *fAM* = fully parametric additive model; *LRh* = log-rank test in hypothetical world; (im)bal = (im)balanced variable, i.e. the groups occur with (un)equal probabilities; bin = binary variable; 4 grps = variable with four groups; X = categorical covariate of interest, D= demographic variables.

Correlation of the p-values: While the sizes of both the log-rank type test and of the semi-parametric additive model are acceptable, the actual p values do not coincide as well. The correlation of the p values of the *LRt* and *sAM* in the above examples in the upper left part of the table is around 0.72. When the proportion of excess hazard deaths becomes large (right part of the table), both tests' results become more similar to their versions in the hypothetical world (which are equal), hence the correlation in all the above examples in the right part of the table is above 0.98. The low correlation implies that one has to make a choice between which test to perform (both *LRt* and *sAM* simultaneously reject the null hypothesis in around 3% of the cases, at least one of the two tests rejects the null hypothesis in 7% of the cases).

Log-rank type test and regression - comparison of the power

Comments on the simulation scenarios: Since we know that both tests simplify to the same test statistic in the hypothetical world, we can expect them to respond to the same alternative hypotheses also with the real world data. This reasoning is checked with simulations reported in Table 4.2. We look at several situations, in particular, we add two situations (last two rows in the table) where the effect of X on excess hazard is not constant in time and hence the proportional excess hazards assumption of the additive model is not met. We simulate crossing hazards, first in a situation where the overall

effect is approximately 0 (β starts at 0.5 and changes to -0.5), and second, in a situation where the overall effect is similar as in other simulations (β starts at 0 and changes to 1).

In all cases, the power of the regression models is expected to be higher than the power of log-rank type test - regression models work with additional assumptions that are true in our scenarios. We are hence more interested in whether the difference between the log-rank type tests and the regression models is similar across different scenarios, if it is not, we could say that the tests are not susceptible to the same alternative hypotheses and hence have a different interpretation.

Table 4.2: Comparison of the log-rank type test and the additive model: power.

	$\frac{d_E}{d} = 0.62$				$\frac{d_E}{d} = 0.93$			
	LRt	*sAM*	*fAM*	*LRh*	*LRt*	*sAM*	*fAM*	*LRh*
$X \in D$, bal, bin (sex)	0.499	0.561	0.552	0.786	0.878	0.874	0.877	0.903
$X \notin D$, bal, bin	0.487	0.578	0.567	0.792	0.877	0.877	0.876	0.903
$X \notin D$, imbal, bin, ef > 0	0.467	0.391	0.359	0.700	0.857	0.793	0.792	0.847
$X \notin D$, imbal, bin, ef < 0	0.537	0.720	0.711	0.864	0.888	0.931	0.933	0.943
$X \notin D$, bal, 4 grps	0.339	0.408	0.385	0.629	0.736	0.743	0.745	0.787
$X \notin D$, bal, bin, NPH, ef ≈ 0	0.052	0.074	0.062	0.047	0.048	0.053	0.051	0.049
$X \notin D$, bal, bin, NPH	0.524	0.510	0.476	0.819	0.880	0.865	0.861	0.912
	$\frac{d_E}{d} = 0.41$							
	LRt	*sAM*	*fAM*	*LRh*				
$X \in D$, bal, bin (sex)	0.224	0.314	0.267	0.649				
$X \notin D$, bal, bin	0.216	0.355	0.315	0.648				
$X \notin D$, imbal, bin, ef > 0	0.199	0.036	0.008	0.522				
$X \notin D$, imbal, bin, ef < 0	0.230	0.506	0.474	0.760				
$X \notin D$, bal, 4 grps	0.152	0.240	0.167	0.473				
$X \notin D$, bal, bin, NPH, ef ≈ 0	0.052	0.090	0.061	0.052				
$X \notin D$, bal, bin, NPH	0.210	0.269	0.204	0.647				

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *LRt* = log-rank type test; *sAM* = semi-parametric additive model; *fAM* = fully parametric additive model; *LRh* = log-rank test in hypothetical world; X = categorical covariate of interest, D= demographic variables; (im)bal = (im)balanced variable, i.e. the groups occur with (un)equal probabilities; bin = binary variable; 4 grps = variable with four groups; ef = variable's effect; NPH = nonproportional effect.

Results: Several results can be read from Table 4.2:

- As expected, the semi-parametric model test has more power than the log-rank type test in most scenarios. This may be at least partly attributed to the additional implicit assumption of the semi-parametric

model (λ_{E_0} is smooth), which is true in our simulation scenarios (λ_{E_0} is taken as a constant in simulations). Interestingly, the power of the fully parametric model is not higher than the power of the semi-parametric model, though it works with the additional information that the baseline hazard is constant throughout the interval.

- With the 41% and 62% cases, the difference between the power of the log-rank type test and the semi-parametric model is similar in all scenarios of proportional hazards and balanced covariates. As the proportion of excess hazard deaths increases, the power of all tests becomes similar.

- Holding other simulation parameters equal, the power changes in the case of imbalanced groups. When the more common group has a lower hazard ($ef < 0$), the power of any test gets higher, the opposite effect on power can be seen when patients in the more common group have a higher hazard. Interestingly, this effect seems more pronounced with the regression models than with the log-rank type test, an explanation for this may be seen in the results of the 41 % scenario: with an extremely low number of events due to excess hazard, the model fitting procedure does not converge, leading to huge variances and unreliable results. The log-rank test thus seems the more stable and reliable option.

- When the proportionality assumption fails, the power of log-rank type test and the regression models becomes very similar, indicating that the tests not only have the same null hypotheses but also follow the same logic which makes them susceptible to the same alternatives. For example, none of methods can detect crossing hazards when the average effect is 0.

Based on these results, we can conclude there is no important difference between the interpretation of the log-rank type test and the test of a coefficient in a univariate additive model, but the power in the individual scenarios may be higher with the models if their additional assumptions are met.

Further notes on the power of log-rank type test

Comments on the simulation scenarios: The excess hazard mortality is a crucial factor for the power of the tests in the hypothetical world,

however, the power in the real world depends also on the proportion of the hypothetical world deaths that we actually observe. The population hazard can thus be seen as a nuisance factor. To illustrate this fact with simulations we consider two scenarios (columns A and B in Table 4.3) with equal distribution of D (and hence equal λ_P values) and equal baseline excess hazard (λ_{E_0}) values. Working with a centered covariate (age) and changing only the sign of its effect ($|\beta_{AGE}|$ remains equal in both scenarios), we get two scenarios with equal power in the hypothetical world. In column A of Table 4.3, the effect of age is in the same direction as in the population mortality tables (older individuals have a higher population and excess mortality hazard), in column B of Table 4.3, the effect works in the opposite direction. We calculate the proportion of individuals who die due to excess hazard ($T_{Ei} < T_{Pi}$) among all individuals who die in the hypothetical world ($T_{Ei} < \tau$) ('Observed proportion of excess hazard deaths' in Table 4.3). We then add a third scenario, where this proportion is held equal, but the age of individuals is lowered and hence the total number of population deaths is lower.
For these simulations, age was considered as the covariate in question as it has the largest effect on population mortality hazard and changing the direction of its effect can thus make an observable difference. We simplified its distribution and considered a binary covariate (55 or 75 years with 50% probability in columns A and B and 53 and 73 years in column C). Therefore, the proportion of excess deaths among all deaths is no longer equal to 62 %. Note that fixing the mortality in the hypothetical world and in the population, scenarios that provide larger differences could not be designed. We thus repeated the simulations 50000 times, to guarantee that the differences are not a consequence of random variation.

Table 4.3: Results of the log-rank type test (LRt) when varying the proportion of excess hazard deaths observed in real world data among all excess hazard deaths in hypothetical world data ('observed proportion').

	A	B	C
Proportion with $T_P < 10$	0.34	0.34	0.29
Proportion with $T_E < 10$ (1- net survival)	0.35	0.35	0.35
Effect of age	$\beta > 0$	$\beta < 0$	$\beta > 0$
Proportion of patients that die in 10 years	0.55	0.58	0.52
Proportion of excess hazard deaths among all deaths	0.51	0.51	0.57
Observed proportion of all excess hazard deaths	0.83	0.86	0.86
Power of LRh	0.934	0.933	0.935
Power of LRt	0.507	0.538	0.572

Results: Table 4.3 confirms the importance of the amount of information lost in the real world, compared to the information available in the hypothetical world. This cannot be directly estimated with the real world data, but the direction of the covariate effect can serve as a guideline. The power in column B of Table 4.3 is higher as the power in column A, since the proportion of individuals who die due to excess hazard in the real world data is higher (86 % of all hypothetical world deaths compared to 83 % in column A). In fact, the columns A and B also differ in the total number of deaths (more deaths in column B), column C is added to prove that the observed difference in power is not due to the total number of deaths - with equal observed proportion of excess hazard deaths, a lower number of individuals dying from other causes increases the power.

Non-stratified and stratified log-rank type test - comparison of the size

We now turn to the comparison of the stratified and non-stratified version. While we need the homogeneity assumption in theory, we would like to evaluate how important this assumption is in practice.

Comments on the simulation scenarios: We simulate scenarios with two covariates affecting the excess hazard and check whether the non-stratified version remains reliable under the null hypothesis. We try scenarios with balanced and imbalanced covariates, different number of strata and consider covariates that are or are not included in the population tables. Since sex is the only categorical variable in population tables, only one scenario with both $X \in D$ and $S \in D$ is considered here (variable S is age; it is categorized into three groups for the stratified version of the test). Further simulations exploring the number of strata used for stratification when both $X \in D$ and $S \in D$ can be found in Table 4.8. Since no important differences can be observed between the scenarios, we further study the power by picking only two of the scenarios in Table 4.6 with different number of strata and vary the size of the effect of S, these results are presented in Table 4.7.

Results: The size of both tests is very close to 0.05 with all the simulations performed, regardless of the number of strata and the size of the effect of S. This gives us confidence that the non-stratified version can be used reliably even if the hazards are non-homogeneous.

Table 4.4: Comparison of the non-stratified and stratified log-rank type test for different covariate types: size.

	$\frac{d_E}{d} = 0.41$		$\frac{d_E}{d} = 0.62$		$\frac{d_E}{d} = 0.93$	
	LRt	*LRt-str*	*LRt*	*LRt-str*	*LRt*	*LRt-str*
$X \in D$ (sex); $S \notin D$, bal, bin	0.051	0.052	0.051	0.051	0.053	0.052
$X \in D$ (sex); $S \notin D$, imbal, bin	0.047	0.049	0.047	0.048	0.048	0.048
$X \in D$ (sex); $S \notin D$, bal, 10 str	0.047	0.047	0.051	0.050	0.050	0.053
$X \in D$ (sex); $S \notin D$, bal, bin, NPH	0.050	0.049	0.051	0.050	0.050	0.050
$X \notin D$, bin; $S \in D$, bal, bin (sex)	0.050	0.049	0.045	0.043	0.048	0.047
$X \in D$ (sex); $S \in D$, (age; 3 str)	0.049	0.049	0.050	0.051	0.051	0.052

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *LRt, LRt-str* = non-stratified and stratified log-rank type test; (im)bal = (im)balanced variable, i.e. the groups occur with unequal probabilities; bin = binary variable; str = strata; NPH = nonproportional effect; X = categorical covariate of interest (balanced); S = stratification covariate; D = demographic variables.

Table 4.5: Comparison of the non-stratified and stratified log-rank type test for different effect sizes: size.

	$\frac{d_E}{d} = 0.41$		$\frac{d_E}{d} = 0.62$		$\frac{d_E}{d} = 0.93$	
	LRt	*LRt-str*	*LRt*	*LRt-str*	*LRt*	*LRt-str*
$X \in D$; $S \notin D$, bal, bin, ef 0	0.044	0.044	0.051	0.049	0.046	0.045
$X \in D$; $S \notin D$, bal, bin, ef 2x	0.049	0.048	0.050	0.049	0.046	0.047
$X \in D$; $S \notin D$, bal, bin, ef 5x	0.052	0.051	0.048	0.046	0.050	0.053
$X \in D$; $S \notin D$, bal, 10str, ef 0	0.052	0.050	0.050	0.049	0.052	0.052
$X \in D$; $S \notin D$, bal, 10str, ef 2x	0.047	0.046	0.048	0.047	0.046	0.050
$X \in D$; $S \notin D$, bal, 10str, ef 5x	0.051	0.051	0.049	0.050	0.047	0.047

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *LRt, LRt-str* = non-stratified and stratified log-rank type test; bal = balanced variable, i.e. the groups occur with equal probabilities; bin = binary variable, str = strata; ef = variable's effect; X = categorical covariate of interest; S = stratification covariate; D = demographic variables. X (sex) is balanced in all cases.

Non-stratified and stratified test - comparison of the power

Comments on the simulation scenarios: We repeat the same scenarios as in the previous subsection, but now with a non-zero effect of X (equal in all simulations). In Table 4.6 the effect sizes of X and S are comparable in size, in Table 4.7 the effect of S is varied.

Results: The power of the stratified test tends to be larger than that of the non-stratified test, but the difference is not really striking (Table 4.6), the differences become important only when the effect of S is large compared to the effect of X (Table 4.7, other distributions of the covariates might bring larger differences). On the other hand, when there is no effect of S, no power seems to be lost by still performing the stratified test.

Table 4.6: Comparison of the non-stratified and stratified log-rank type test for different covariate types: power.

	$\frac{d_E}{d} = 0.41$		$\frac{d_E}{d} = 0.62$		$\frac{d_E}{d} = 0.93$	
	LRt	*LRt-str*	*LRt*	*LRt-str*	*LRt*	*LRt-str*
$X \in D$ (sex); $S \notin D$, bal, bin	0.215	0.217	0.498	0.501	0.879	0.882
$X \in D$ (sex); $S \notin D$, imbal, bin	0.221	0.222	0.508	0.510	0.872	0.874
$X \in D$ (sex); $S \notin D$, bal, 10 str	0.228	0.230	0.501	0.498	0.875	0.878
$X \in D$ (sex); $S \notin D$, bal, bin, NPH	0.224	0.227	0.514	0.520	0.887	0.893
$X \notin D$, bin; $S \in D$, bal, bin (sex)	0.221	0.222	0.483	0.489	0.872	0.875
$X \in D$ (sex); $S \in D$, (age; 3 str)	0.218	0.221	0.512	0.513	0.871	0.873

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *LRt, LRt-str* = non-stratified and stratified log-rank type test; (im)bal = (im)balanced variable, i.e. the groups occur with unequal probabilities; bin = binary variable; str = strata; NPH = nonproportional effect; X = categorical covariate of interest (balanced); S = stratification covariate; D = demographic variables.

Table 4.7: Comparison of the non-stratified and stratified log-rank type test for different effect sizes: power.

	$\frac{d_E}{d} = 0.41$		$\frac{d_E}{d} = 0.62$		$\frac{d_E}{d} = 0.93$	
	LRt	*LRt-str*	*LRt*	*LRt-str*	*LRt*	*LRt-str*
$X \in D$; $S \notin D$, bal, bin, ef 0	0.224	0.224	0.505	0.505	0.870	0.870
$X \in D$; $S \notin D$, bal, bin, ef 2x	0.235	0.237	0.502	0.519	0.863	0.880
$X \in D$; $S \notin D$, bal, bin, ef 5x	0.235	0.264	0.465	0.559	0.778	0.894
$X \in D$; $S \notin D$, bal, 10str, ef 0	0.223	0.220	0.495	0.496	0.879	0.875
$X \in D$; $S \notin D$, bal, 10str, ef 2x	0.223	0.229	0.486	0.499	0.862	0.878
$X \in D$; $S \notin D$, bal, 10str, ef 5x	0.245	0.271	0.472	0.559	0.772	0.895

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; *LRt, LRt-str* = non-stratified and stratified log-rank type test; bal = balanced variable, i.e. the groups occur with equal probabilities; bin = binary variable, str = strata; ef = variable's effect; X = categorical covariate of interest; S = stratification covariate; D = demographic variables. X (sex) is balanced in all cases.

Stratified test - the effect of the number of strata

Comments on the simulation scenarios: As the last point, we further check the performance of the stratified test in case of many strata. To mimic a real life situation, we compare groups with respect to sex and stratify by age, which we can always expect to be an important factor. We simulate age as a continuous variable with a linear effect on log excess hazard, but then categorize it to allow for stratification. The effect of age is substantially larger than the effect of sex (5 times higher), so that some differences in power can be observed.

Results: Table 4.8 once again confirms that the size of the non-stratified version of the test is appropriate and that the stratified test has more power.

Table 4.8: Comparison of the non-stratified and stratified log-rank type test for different number of strata.

length of age interval for stratification (no. of strata)	$\frac{d_E}{d} = 0.41$ size	power	$\frac{d_E}{d} = 0.62$ size	power
non-stratified	0.047	0.226	0.042	0.500
10 years (3 strata)	0.051	0.248	0.049	0.538
5 years (6 strata)	0.051	0.246	0.049	0.546
1 year (30 strata)	0.054	0.240	0.051	0.534
6 months (60 strata)	0.051	0.233	0.050	0.516
1 month (360 strata)	0.049	0.174	0.050	0.368

d = number of deaths from any cause; d_E = number of deaths due to excess hazard; $X \in D$ (sex is the categorical covariate of interest), $S \in D$ (age (grouped) is the stratification covariate).

However, age is a continuous variable and thus the question is, how many strata to make. We can see that in the '62% case' the power is best with 6 strata, but not much worse with only 3 strata, which is rather few considering that a strong effect of age was simulated. On the other hand, splitting to 30 strata which leaves some strata with only few events (practically all simulation runs include strata with less than 5 events), still provides an improved power compared to the non-stratified version. However, if the strata are far too many (last row of the Table 4.8), the power is importantly decreased. The reason is that many strata are without events or there is only one group within stratum and hence the information on excess hazard carried by the patients in these strata is not included in the estimation (on average a third of the patients belong to such strata). Similar behaviour can be observed in the '41% case', except for the fact that the power is highest with only 3 strata.

4.1.4 Example

Let us now focus on the data on myocardial infarction. Out of 494 patients, 204 died (0.41). To get some idea of the power we can expect with our sample, we consider the proportion expected to die due to excess hazard (0.22; PP estimator) and the proportion expected to die in the population (0.31; population tables). The effect of sex in our sample is in the opposite direction as in the population, which, judging from the simulations, is also a positive indicator for the power. Figure 4.1 presents the net survival curves estimated by the PP estimator, we observe a marked difference between men and women that is confirmed by the log-rank type test with respect to sex ($p = 0.02$).

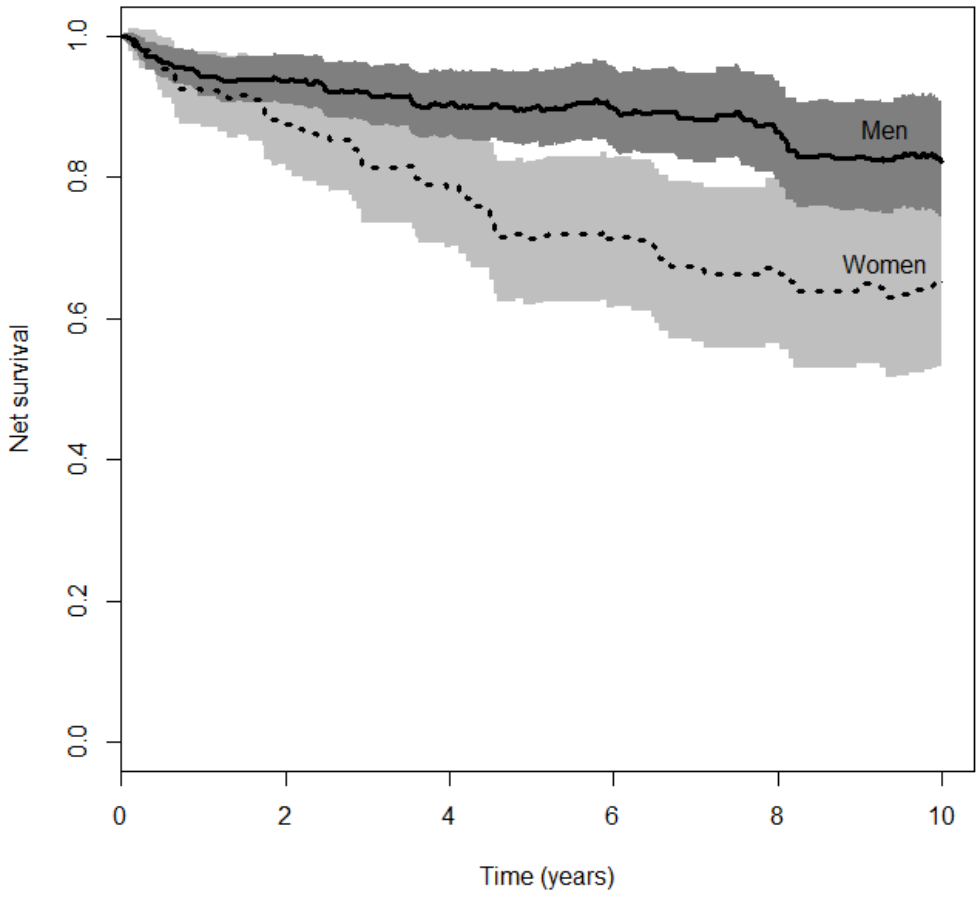

Figure 4.1: Comparison of PP estimator for men and women.

To confirm our simulation results, we check also the semi-parametric and the fully parametric model (with one parameter for the baseline hazard in the first year of follow-up and another afterwards). Both yield equal interpretation, with p values 0.001 and 0.006, respectively.
Of course, both men and women differ in age (age span 45-75) and a univariate model including age shows that age is an important covariate in terms of excess hazard. If interested in the effect of sex conditional on age being the same, we can consider the stratified model. However, the stratification with respect to age cannot be very fine, a yearly stratification would imply that 10% of the patients (mostly the young ones) are omitted from the

calculations.

Following our simulation results, we therefore limit ourselves to 2-year or 5-year strata, which results in p values equal to 0.18 and 0.09, respectively.

The different result than in the non-stratified test is expected as age and sex are not independent (women are on average 4.8 years older) and the usage of the stratified test thus implies a different interpretation. If we include age into a multivariate model and thus assume linearity of the effect, we get a significant effect of sex in the case of the semi-parametric model ($p = 0.03$) and a borderline significant effect in the case of the fully parametric model ($p = 0.058$). Since the linearity of age is hard to judge on our data (especially with the younger patients, where there are only a few individuals), the two models seem rather unreliable.

We can therefore conclude that the net survival curves differ significantly by sex, the non-stratified log-rank type test as a very reliable option can be used to show this. Whether or not this difference can be fully explained by the different age at infarction or persists within patients of same age, remains a question that is hard to respond to reliably with our sample, since the age-distribution is too wide for such a small sample and thus very little can be said about the youngest patients.

Chapter 5

Conclusions

It is the time you have wasted for your rose that makes your rose so important.
Antoine de Saint-Exupéry, The Little Prince

Our primary focus has been estimation and comparison of net survival, i.e. a quantity which is independent of λ_P and represents the survival of cancer patients in a hypothetical world where they cannot die from other causes. This interpretation is usually quite unrealistic (at least for events that preclude other events from happening) and relies on an assumption which cannot be tested from the data.

We have soon realized that the whole field of relative survival analysis operates with a lot of assumptions, some of them are difficult to justify. We have explored and described them thoroughly. Some assumptions are really not necessary and are there just to ensure nice interpretations (or to define measures that suppose to have nice interpretations, e.g. independence of latent times in the definition of net survival) but others are a consequence of the available data (e.g. population mortality tables are used to represent the hazard for death from other causes and they are not split with respect to important covariates such as smoking).

We have attempted to minimize the number of assumptions required to report a measure of cancer patients survival and in particular, we have tried to avoid the assumptions required to interpret net survival as a survival in a hypothetical world. We have defined a new measure called marginal relative survival. It is derived from a generalization of a model that is used to define net survival, it has less assumptions and therefore different interpretation

(which exists in the real world). However, it simplifies to net survival when its assumptions are valid.

Marginal relative survival compares patients to their counterparts in the population who are defined by some demographic covariates. This measure can be always estimated and interpreted as the average of ratios. However, it may depend on hazard of dying from other causes. This will occur if the population mortality tables are not split with respect to all the covariates that have an effect on it. As such, this dependence is a consequence of the available data and there is no simple way around it.

We believe that the main step forward of this work represents the recognition of all the underlying assumptions of net survival and clarification of the effects they have on the estimation and interpretation of this measure. We have realized that there is a distinction between the ideal that researchers from cancer registries would like to estimate (i.e. net survival) but is out of reach and the quantity that is closest to this ideal and can really be estimated. We have defined marginal relative survival to emphasize this difference.

We have proposed a new pseudo based approach to estimation of marginal relative survival. The new estimator is constructed directly from the definition of marginal relative survival by using pseudo observations and merging them with population mortality tables. We have derived the variance of the new estimator and proposed two estimators for it: the precise and the approximative one. We have mimicked the derivation of the estimator of variance of Kaplan-Meier estimator to obtain these two estimators. We have explored the properties of the new estimator by simulations. It gives estimates that are practically identical to PP estimates. The precise formula for the variance estimates the variance well (i.e. it gives coverages close to the nominal level) and the approximative formula slightly underestimates it.

We have further extended the pseudo based approach to data with discretely measured time. In this case we have used the standard lifetable estimator instead of Kaplan-Meier estimator to calculate pseudo observations. We have plugged them into the definition of marginal relative survival estimator and have modified the variance estimators. We have used simulations to show that the newly proposed estimation procedure outperforms the PP estimator in terms of bias, especially when the aggregation time window is wide. Furthermore, the modified variance estimators seem to work reasonably well.

The last topic of this thesis is the comparison of net and marginal relative survival between different groups. A log-rank type test has been proposed by

Grafféo et al. [13] and we have implemented it into the `R` package `relsurv` [24, 22] to ensure its availability and usability. We have used `C` code to improve efficiency and to enable its use on large data sets from cancer registries that are common in this field. We have built upon the known equality of the log-rank statistic and the score statistic from the Cox model and have been looking for a similar relationship in the field of relative survival where the log-rank type test can be compared to the additive semi-parametric model. The additive form of the model prevents the population part from cancelling out in the likelihood and there is no theoretical equality between the two statistics. We have found out that in each particular situation different conclusions can be drawn based on the chosen method. We have therefore explored if the tests behave in a similar manner and to what extent.

Simulation results show that the sizes of both the log-rank type test and the semi-parametric additive model are close to the nominal level but the p values do not coincide as well. The correlation of the p values decreases with decreasing proportion of deaths due to excess hazard among all deaths. The semi-parametric additive model test is slightly more powerful in some of the scenarios but this may be partially attributed to its implicit assumption about smooth baseline. Nevertheless, the differences in power seem to be similar in all cases of proportional hazards and balanced covariates. When the proportionality assumption is violated, both tests lose power. This leads us to say that both tests can be interpreted in the same way but the log-rank type test seems to be more stable and simpler alternative.

The other issue we have wished to clarify was the role of the stratified version of the test. In the paper by Grafféo et al. [13] the stratified version is used to take into account the inhomogeneneity of the baseline hazard between different groups. We have compared the behaviour of the stratified and non-stratified version. Our simulations reveal that both the stratified and non-stratified version can be used reliably even if the hazards are non-homogeneous. The power of the stratified test with a reasonable number of strata tends to be larger when compared to the power of the non-stratified test but the differences are usually small. Note however that the stratified and non-stratified version have different interpretation.

We have recognized pseudo observations as a very valuable tool that has helped us define the estimator of the new measure but that is not all they offer. They present a new starting point that has enabled us to construct estimators on survival scale which are simpler to see through and whose properties are therefore easier to grasp. We already see new extensions to

regression modelling where the ratios that were used in the estimator of marginal relative survival could be used as responses in a generalized linear model. This would allow us to explore the effect of a certain covariate on marginal relative survival and represents a potential challenge for the future.

Appendix A

Povzetek dela

A.1 Uvod

Analiza relativnega preživetja je podpodročje analize preživetja, ki se ukvarja s sotveganji na podatkih, kjer vzrok smrti ni znan ali zanesljiv. Najpogosteje se uporablja pri analizi preživetja bolnikov z rakom, ki lahko umrejo zaradi raka ali pa iz drugih vzrokov. Manjkajočo informacijo o vzroku smrti nadomestimo z vpeljavo podatkov iz populacijskih tabel umrljivosti.

Populacijske tabele umrljivosti (tudi tablice umrljivosti) vsebujejo informacijo o številu posameznikov neke starosti in spola, ki so v danem letu umrli. Z njihovo pomočjo lahko izračunamo kakšno tveganje za smrt ima oseba določene starosti in spola v določenem letu. Če imamo opravka z rakom, ki se v populaciji pojavi redko, ta rak nima velikega vpliva na populacijske tabele umrljivosti in jih zato lahko uporabimo kot vir informacij o tveganju za smrt iz drugih vzrokov.

V analizi relativnega preživetja se uporablja več mer, med njimi so pričakovano preživetje (*ang. expected survival*), razmerje preživetij (*ang. relative survival ratio*), lahko pa ocenimo tudi kumulativno funkcijo pojavnosti (*ang. cumulative incidence function*). Vsaka izmed njih povzame informacijo o preživetju bolnikov z rakom nekoliko drugače. V tem delu se bomo osredotočili na čisto preživetje, ker je to po definiciji neodvisno od tveganja za smrt iz drugih vzrokov in naj bi kot tako omogočalo primerjavo preživetja bolnikov z rakom med različnimi populacijami.

A.2 Čisto in robno relativno preživetje

V analizi relativnega preživetja je običajno izhodiščna točka aditivni model

$$\lambda_O(t|X) = \lambda_E(t|X) + \lambda_P(t|X),$$

kjer je $\lambda_O(t|X)$ skupno tveganje, količini $\lambda_E(t|X)$ in $\lambda_P(t|X)$ pa sta definirani kot tveganje za smrt zaradi raka in zaradi drugih (populacijskih) vzrokov. Pri tem pa pogosto ni jasno specificirano ali sta ti dve količini sotveganji ali gre za robni tveganji. To lahko vodi do mnogih napačnih razlag in nerazumevanja mer, ki jih ocenjujemo. V ta namen bomo najprej jasno pojasnili razliko med tema dvema vrstama tveganj, nato pa pravilno definirali čisto preživetje in pojasnili pomanjkljivosti njegove definicije.

Naj bo T_E latentni čas do smrti zaradi raka in T_P latentni čas do smrti zaradi drugih vzrokov. S $T = \min(T_E, T_P)$ označimo čas do dogodka (smrti), s $\widetilde{T} = \min(T, C)$ pa opaženi čas, kjer C predstavlja čas do krnjenja. Naj bosta δ_E in δ_P indikatorja smrti zaradi raka in iz drugih vzrokov.

Sotveganji sta definirani kot

$$\lambda_E^*(t|X) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \le T < t + \Delta t, \delta_E | T \ge t, X),$$
$$\lambda_P^*(t|D) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \le T < t + \Delta t, \delta_P | T \ge t, D),$$

robni tveganji pa kot

$$\lambda_E(t|X) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \le T_E < t + \Delta t | T_E \ge t, X),$$
$$\lambda_P(t|D) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \le T_P < t + \Delta t | T_P \ge t, D).$$

Opazimo lahko, da v definiciji sotveganja nastopa le čas do dogodka, pri robnih tveganjih pa predpostavljamo obstoj latentnih časov do smrti zaradi danih vzrokov. Če latentna časa obstajata, se pogoj $\{t \le T < t + \Delta t, \delta_E\}$ v definiciji sotveganja $\lambda_E^*(t|X)$ poenostavi v $\{t \le T_E < t + \Delta t\}$. Podobno velja za $\lambda_P^*(t|X)$.

Izhodiščni model v analizi relativnega preživetja je tako model $\lambda_O = \lambda_E^* + \lambda_P^*$, za definicijo čistega preživetja pa privzamemo, da je ta model dejansko enak $\lambda_O = \lambda_E + \lambda_P$. Čisto preživetje definiramo kot $S_E(t) = \exp\Big(-\int_0^t \lambda_E(u)du\Big)$,

pri tem pa predpostavljamo, da je $\lambda_E^* = \lambda_E$ in $\lambda_P^* = \lambda_P$. V literaturi se kot pogoj, ki zagotavlja to enakost, pogosto pojavi pogoj o neodvisnosti latentnih časov do smrti T_E in T_P. Ta pogoj je sicer res zadosten, ni pa nujno potreben. Kljub temu pa je precej nerealistično pričakovati, da so sotveganja kar enaka robnim tveganjem. Ta problem je precej povezan z drugo pomanjkljivostjo čistega preživetja. S tem, ko predpostavimo, da je sotveganje kar enako robnemu tveganju, namreč pridobimo možnost interpretiranja čistega preživetja kot funkcije preživetja. Ker pa obravnavamo le tveganje za smrt zaradi raka, jo lahko interpretiramo le kot preživetje v hipotetičnem svetu, kjer bolniki lahko umrejo le zaradi raka. Ta svet pa seveda ne obstaja in ga ne moremo opazovati. Če torej povzamemo: predpostavko, da so sotveganja enaka robnim tveganjim, je težko utemeljiti in nemogoče preveriti na podatkih, nudi pa nam lepo interpretacijo v smislu funkcije preživetja, vedar le v hipotetičnem svetu.

To nas je napeljalo k temu, da bi poiskali mero, ki ne bo temeljila na tej neprevervljivi predpostavki. Začnemo lahko precej splošno s posplošitvijo modela sotveganj. Naj bo

$$\lambda_O(t|X) = \widetilde{\lambda}_E(t|X) + \widetilde{\lambda}_P(t|D),$$

kjer $\widetilde{\lambda}_E(t|X)$ predstavlja prispevek raka, ki ni nujno pozitiven (torej ni nujno tveganje), $\widetilde{\lambda}_P(t|D)$ pa predstavlja tveganje za smrt zdravega vrstnika z istimi demografskimi kovariatami D. Ta model transformiramo s funkcionalom $x \mapsto \exp(-\int_0^t x(u)du)$ in dobimo

$$\widetilde{S}_E(t|X) = \frac{S_O(t|X)}{\widetilde{S}_P(t|D)},$$

kjer je $S_O(t|X)$ skupno preživetje, $\widetilde{S}_P(t|D)$ pa preživetje zdravega vrstnika z istimi demografskimi kovariatami. Količini $\widetilde{S}_E(t|X)$ pravimo *individualno relativno preživetje*, za skupino pa definiramo *robno relativno preživetje*

$$\widetilde{S}_E(t) = \int \widetilde{S}_E(t|x)dH(x),$$

kjer je H porazdelitvena funkcija X. Cena za splošnost v definiciji $\widetilde{S}_E(t)$ je dejstvo, da to ni nujno funkcija preživetja, jo pa lahko nedvoumno interpretiramo kot povprečje razmerij med skupnim preživetjem in preživetjem vrstnika. Čisto preživetje še vedno lahko obravnavamo kot poseben primer robnega relativnega preživetja, tj. ko so predpostavke, ki so potrebne za definicijo čistega preživetja, izpolnjene, lahko robno relativno preživetje interpretiramo kot čisto preživetje.

A.3 Ocenjevanje robnega relativnega preživetja

Cenilko za ocenjevanje čistega preživetja bomo skonstruirali direktno iz definicije mere. Na vzorcu velikosti n je cenilka definirana kot

$$\widehat{\widetilde{S}}_E(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{S}_{Oi}(t)}{\widetilde{S}_{Pi}(t)},$$

kjer je $\widetilde{S}_{Pi}(t)$ preživetje zdravega vrstnika z istimi demografskimi kovariatami, $\widehat{S}_{Oi}$ pa je cenilka individualnega skupnega preživetja. Za slednjo lahko v primeru, ko je edino krnjenje v podatkih tisto na koncu opazovanja, preprosto uporabimo indikatorsko funkcijo $I(T_i > t)$. Ta pristop ne deluje v primeru, ko imamo v podatkih posameznike, ki so krnjeni pred koncem opazovanja (kar je običajno v praksi). V tem primeru predlagamo uporabo psevdo vrednosti, ki jih lahko definiramo vedno, ko imamo na voljo cenilko (dovolj lepih lastnosti), ki oceni mero na celem vzorcu. Pri tem je $i$ta *psevdo vrednost* definirana kot

$$\widehat{S}_{Oi}(t) = n \cdot \widehat{S}_O(t) - (n-1) \cdot \widehat{S}_O^{(-i)}(t),$$

kjer je $\widehat{S}_O(t)$ cenilka uporabljena na celotnem vzorcu, $\widehat{S}_O^{(-i)}(t)$ pa je ista cenilka uporabljena na reduciranem vzorcu, ki ga dobimo iz celotnega z odstranitvijo vrednosti i-tega pacienta. Prednost psevdo vrednosti je, da so definirane za vse paciente (tudi krnjene) za vse čase do največjega opaženega časa, pri tem pa $i$ta psevdo vrednost predstavlja prispevek $i$tega pacienta k oceni skupnega preživetja. V našem primeru za oceno skupnega preživetja v izračunu psevdo vrednosti uporabljamo cenilko KM (Kaplan-Meier). Omeniti velja še, da psevdo vrednosti lahko izračunamo tudi za druge količine (na primer za kumulativno funkcijo pojavnosti).

Za cenilko robnega relativnega preživetja, ki uporablja psevdo vrednosti izračunane na podlagi cenilke KM, smo izračunali varianco in zanjo predlagali dve cenilki. Prva upošteva odvisnost med psevdo vrednostmi in je zato bolj natančna in računsko zahtevna. Druga pa uporabi dejstvo, da $i$ta psevdo vrednost asimptotsko temelji le na informaciji, ki jo nosi $i$ti pacient in zato asimptotsko ni odvisna od drugih. Ob predpostavki, da to velja tudi na končnih vzorcih smo izpeljali drug pristop k ocenjevanju variance, ki je računsko enostavnejši.

Pri konstrukciji teh dveh cenilk smo uporabili teorijo procesov štetja in martingalov, zgledovali pa smo se po izpeljavi cenilke za varianco cenilke KM.

Novo cenilko relativnega robnega preživetja smo primerjali tudi s cenilko PP, ki ocenjuje čisto preživetje. Cenilki dajeta praktično enake rezultate, kar smo pokazali v teoriji in potrdili s simulacijami. Slednje smo uporabili tudi za preverjanje lastnosti cenilk za ocenjevanje variance. Ugotovili smo, da natančna cenilka dobro oceni varianco, kar je vidno tudi pri pokritjih, ki so blizu nominalne vrednosti. Aproksimativna cenilka variance rahlo podceni varianco, vendar je napaka v večini realističnih primerov majhna; pokritje intervalov zaupanja je v večini primerov le rahlo odstopalo od nominalne vrednosti.

Cenilka PP, ki ocenjuje čisto preživetje, je bila razvita za zvezno merjen čas in daje pristranske rezultate, če čas ni merjen zvezno. To pomeni, da za paciente ne poznamo natančnega časa smrti temveč vemo le, da so umrli znotraj nekega intervala (poznamo le teden ali mesec smrti, ne pa dneva). Cenilko relativnega robnega preživetja lahko modificiramo za uporabo na podatkih takega tipa. Kadar želimo oceniti preživetje ob nekem času, pri čemer čas ni merjen zvezno temveč le diskretno, problem predstavljajo krnjeni posamezniki. Za njih namreč vemo, da so bili krnjeni nekje znotraj intervala med točkama, ob katerih imamo meritve. To je potrebno pravilno upoštevati pri ocenjevanju. Cenilka KM, ki smo jo uporabljali pri izračunu psevdo vrednosti v primeru zvezno merjenega časa, krnjene obravnava kot da so prisotni celoten interval (do časa, ko imamo meritev). Namesto nje bomo zato uporabili t.i. cenilko tablic preživetja. Ta se od cenilke KM razlikuje v tem, da uporabi ocenjeno efektivno število ljudi, ki so še izpostavljeni tveganju za smrt. To doseže tako, da število ljudi, ki so še izpostavljeni tveganju na začetku nekega intervala zmanjša za nek delež krnjenih v tistem intervalu - s tem dosežemo, da krnjeni niso izpostavljeni tveganju celoten interval. Delež krnjenih, ki se najpogosteje uporabi, je 0.5.

Psevdo vrednosti v cenilki robnega relativnega preživetja bomo zato izračunali na podlagi cenilke tablic preživetja. S simulacijami smo preverili obnašanje dobljene cenilke. Ima manjšo pristranskost od cenilke PP: to je še posebej očitno, ko so intervali, v katerih je merjen čas, zelo široki in je delež krnjenih posameznikov visok. Prilagodili smo tudi formulo za oceno variance in pogledali, kakšna pokritja dosežemo s to cenilko. Dobimo pokritja, ki so zelo blizu nominalni vrednosti. Zaključimo lahko, da na podatkih z diskretno merjenim časom nova cenilka deluje bolje kot cenilka PP.

A.4 Primerjava krivulj čistega preživetja

V četrtem poglavju tega dela smo se posvetili primerjavi čistega in robnega relativnega preživetja med skupinami. Grafféo et al. [13] so predlagali test oblike log-rank. Ta test združuje idejo testa log-rank z idejo uteževanja iz cenilke PP. Skonstruiran je bil za primerjavo čistega preživetja med skupinami definiranimi z neko kategorialno spremenljivko. Ker pa smo ugotovili, da ocenjene vrednosti čistega preživetja praktično sovpadajo z ocenami robnega relativnega preživetja, lahko rečemo, da test primerja robno relativno preživetje med skupinami.

Alternativa testu oblike log-rank so regresijski modeli. Pošteno primerjavo omogočajo le testi, ki temeljijo na modelu

$$\lambda_O(t|X) = \tilde{\lambda}_P(t|D) + \tilde{\lambda}_{E_0}(t)e^{\beta X},$$

kjer $\tilde{\lambda}_{E_0}(t)$ ni specificiran, saj tudi test oblike log-rank ne dela predpostavk o porazdelitvi. Tak regresijski model je na voljo v paketu `relsurv` [24, 22], zato smo test oblike log-rank sprogramirali in vključili v ta paket. Nato smo se osredotočili na primerjavo testa koeficienta iz regresijskega modela s testom oblike log-rank. Pri tem nas je vodila enakost testa log-rank in testa na podlagi zbira za koeficient iz Coxovega modela. Hitro smo ugotovili, da testni statistiki (in zato rezultati testov) v našem primeru nista identični. Krivec za to je populacijski del v regresijskem modelu, ki se v funkciji delnega verjetja ne pokrajša.

Preveriti smo želeli, če se testa obnašata podobno. V ta namen smo izvedli simulacijsko študijo. Najprej smo preverili, da imata testa primerno velikost, nato pa smo primerjali moči pri različnih alternativnih hipotezah. Razmerje moči je podobno pri različnih alternativnih hipotezah, prav tako pa sta oba testa neobčutljiva za nesorazmerna tveganja. Preverili smo še obnašanje stratificirane verzije testa oblike log-rank. Ugotovili smo, da lahko nestratificirani test uporabljamo zanesljivo, čeprav tveganje znotraj skupin ni homogeno, ocenjena velikost testa je namreč sovpadala z nominalno vrednostjo. Ima pa v teh primerih stratificirani test rahlo večjo moč, vendar razlika ni velika. Opazili smo še, da stratificirani test izgubi moč, če uporabimo preveliko število stratumov (kar pomeni, da so stratumi brez dogodkov).

A.5 Zaključek

Naslov teze se osredotoča ne ocenjevanje in primerjavo čistega preživetja. Tekom raziskovanja lastnosti te mere in njenih cenilk, smo ugotovili, da se znotraj področja analize relativnega preživetja uporablja veliko predpostavk. Nekatere izmed njih so posledica podatkov, ki jih imamo na voljo (predpostavka, da tveganje za smrt iz drugih vzrokov lahko razberemo iz populacijskih tabel umrljivosti), druge pa so umetno ustvarjene z namenom, da dobimo mere z lepo interpretacijo. Primer tega je predpostavka, da je sotveganje enako robnemu tveganju, ki jo uporabimo za definicijo čistega preživetja. To predpostavko je težko utemeljiti, na podatkih se je ne da testirati, zato je interpretacija mere vprašljiva. Čisto preživetje zaradi te predpostavke res lahko interpretiramo kot preživetje, vendar le v hipotetičnem svetu, kjer ljudje lahko umrejo le zaradi raka.

Da bi se izognili nerealnim predpostavkam čistega preživetja, smo definirali novo mero, imenovano robno relativno preživetje. Ta prav tako kot čisto preživetje izvira iz aditivnega modela, vendar so sumandi v njem splošnejše funkcije. Cena za to splošnost je dejstvo, da dobljena mera v splošnem ni funkcija preživetja.

Predlagali smo cenilko za robno relativno preživetje, ki je skonstrirana direktno iz definicije mere in uporablja psevdo vrednosti. To nam je omogočilo njeno razširitev na podatke z diskretno merjenim časom. Na podatkih z zvezno merjenim časom se ocene dobljene na podlagi nove cenilke in cenilke PP (ki je bila skonstrirana za ocenjevanje čistega preživetja) praktično ujemajo (kar smo preverili v teoriji in s simulacijami), na podatkih z diskretno merjenim časom pa nova cenilka daje manj pristranske rezultate kot cenilka PP. Za novo cenilko smo izračunali varianco in predlagali dve cenilki za njeno ocenjevanje: natančno in aproksimativno, ki je tudi hitrejša. S simulacijami smo preverili tudi, kakšno pokritje dobimo z danima cenilkama. Natančna cenilka daje pokritja, ki so zelo blizu nominalni vrednosti, aproksimativna cenilka pa rahlo podceni varianco vendar so razlike v realističnih scenarijih majhne.

Na koncu smo se osredotočili še na primerjavo krivulj čistega in robnega relativnega preživetja med skupinami. Želeli smo potegniti vzporednice z analizo preživetja, kjer je znano, da sta log-rank test in test koeficienta iz Coxovega modela na podlagi zbira identična. Primerjali smo test oblike log-rank in teste koeficientov iz aditivnega regresijskega modela, v katerem osnovno tveganje sumanda, ki pripada tveganju zaradi raka, ni specificirano. Ugotovili

smo, da testa nista identična, se pa obnašata podobno. Odzivata se na iste alternativne hipoteze, prav tako pa sta oba neobčutljiva za alternativne hipoteze pri katerih efekt spremeni predznak. Pokazali smo še, da lahko nestratificirano verzijo testa uporabimo zanesljivo tudi v primeru, ko tveganje po skupinah ni homogeno. Ima pa stratificirani test v takih primerih rahlo večjo moč, če število dogodkov znotraj stratumov ni premajhno.

Naše mnenje je, da robno relativno preživetje predtavlja majhen, a pomemben korak naprej v analizi relativnega preživetja. Enostavnost definicije te mere ponuja boljše razumevanje pasti interpretacije čistega preživetja kot funkcije preživetja.

Robno relativno preživetje primerja paciente z njihovimi vrstniki, ki so definirani prek demografskih kovariat. Ta primerjava je lahko zavajujoča, če nam manjka pomembna kovariata kot je kajenje. Treba je poudariti, da to ni pomanjkljivost mere ampak posledica pomanjkljivih podatkov, ki jih imamo na voljo.

Psevdo vrednosti so se izkazale kot zelo široko uporabno orodje, ki nam je omogočilo konstrukcijo cenilke neposredno iz definicije mere. Naravno pa se je pojavila njihova naslednja možna uporaba, saj bi lahko kvociente, ki smo jih uporabili v cenilki nove mere, uporabili kot izide v regresijskih modelih s katerimi bi lahko raziskali vpliv kovariat na robno relativno preživetje. To predstavlja izziv za prihodnost.

Bibliography

[1] (2016). Human mortality database. Available at www.mortality.org or www.humanmortality.de (data downloaded on 10.3.2016).

[2] Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

[3] Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12):1074–1088.

[4] Andersen, P. K., Klein, J. P., and Rosthøj, S. (2003). Generalized linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27.

[5] Andersen, P. K. and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99.

[6] Corazziari, I., Quinn, M., and Capocaccia, R. (2004). Standard cancer patient population for age standardizing survival ratios. *European Journal of Cancer*, 40(15):2307–2316.

[7] Danieli, C., Remontet, L., Bossard, N., Roche, L., and Belot, A. (2012). Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine*, 31(8):775–786.

[8] De Angelis, R., Francisci, S., Baili, P., Marchesi, F., Roazzi, P., Belot, A., Crocetti, E., Pury, P., Knijn, A., Coleman, M., and Capocaccia, R. (2009). The EUROCARE-4 database on cancer survival in Europe: data standardisation, quality control and methods of statistical analysis. *European Journal of Cancer*, 45(6):909–930.

[9] Dickman, P. W., Lambert, P. C., Coviello, E., and Rutherford, M. J. (2013). Estimating net survival in population-based cancer studies, letter to editor. *International Journal of Cancer*, 133(2):519–521.

[10] Estève, J., Benhamou, E., Croasdale, M., and Raymond, M. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, 9(5):529–538.

[11] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. John Wiley and Sons, Inc.

[12] Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Estève, J., Gouvernet, J., and Faivre, J. (2003). A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine*, 22(17):2767–84.

[13] Grafféo, N., Castell, F., Belot, A., and Giorgi, R. (2016). A log-rank-type test to compare net survival distributions. *Biometrics*, 72(3):760–769.

[14] Grafféo, N., Jooste, V., and Giorgi, R. (2012). The impact of additional life-table variables on excess mortality estimates. *Statistics in Medicine*, 31(30):4219–4230.

[15] Graw, F., Gerds, T., and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255.

[16] Jacobsen, M. and Martinussen, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862.

[17] Kalbfleisch, J. D. and Prentice, R. (2002). *The statistical analysis of failure time data (2nd edition)*. John Wiley and Sons, Inc., Hoboken, NJ.

[18] Lambert, P. C., Dickman, P. W., and Rutherford, M. J. (2015). Comparison of different approaches to estimating age standardized net survival. *BMC Medical Research Methodology*, 15(1). URL: http://www.biomedcentral.com/content/pdf/s12874-015-0057-3.pdf.

[19] Lambert, P. C., Smith, L. K., Jones, R. J., and Botha, J. L. (2005). Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine*, 24(24):3871–3885.

[20] Lawless, J. (1982). *Statistical models and methods for lifetime data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

[21] Mahboubi, A., Abrahamowicz, M., Giorgi, R., Binquet, C., Bonithon-Kopp, C., and Quantin, C. (2011). Flexible modeling of the effects of continuous prognostic factors in relative survival. *Statistics in Medicine*, 30(12):1351–1365.

[22] Pavlič, K. and Pohar Perme, M. (2017). On comparison of net survival curves. *BMC Medical Research Methodology*, 17(1):79.

[23] Pohar, M. and Stare, J. (2006). Relative survival analysis in R. *Computer Methods and Programs in Biomedicine*, 81(3):272–278.

[24] Pohar Perme, M. (2017). *relsurv: Relative Survival*. R package version 2.1-1.

[25] Pohar Perme, M., Henderson, R., and Stare, J. (2009). An approach to estimation in relative survival regression. *Biostatistics*, 10(1):136–146.

[26] Pohar Perme, M., Stare, J., and Estève, J. (2012). On estimation in relative survival. *Biometrics*, 68(1):113–120.

[27] Pokhrel, A. and Hakulinen, T. (2009). Age-standardisation of relative survival ratios of cancer patients in a comparison between countries, genders and time periods. *European Journal of Cancer*, 45(4):642–647.

[28] Rebolj, A. (2016). *Nepristranske cenilke v relativnem preživetju: doktorsko delo*. PhD thesis, University of Ljubljana, Ljubljana, Slovenia.

[29] Rice, J. A. (2009). *Mathematical Statistics and Data Analysis*. Duxbury advanced series, Cengage learning.

[30] Roche, L., Danieli, C., Belot, A., Grosclaude, P., Bouvier, A. M., Velten, M., Iwaz, J., Remontet, L., and Bossard, N. (2012). Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. *International Journal of Cancer*, 132(10):2359–69.

[31] Sasieni, P. and Brentnall, A. R. (2016). On standardized relative survival. *Biometrics*, 73(2):473–482.

[32] Sasieni, P. D. (1996). Proportional excess hazards. *Biometrika*, 83(1):127–141.

[33] Seppä, K., Hakulinen, T., Läärä, E., and Pitkäniemi, J. (2016). Comparing net survival estimators of cancer patients. *Statistics in Medicine*, 35(11):1866–1879.

[34] Seppä, K., Hakulinen, T., and Pokhrel, A. (2015). Choosing the net survival method for cancer survival estimation. *European Journal of Cancer*, 51(9):1123 – 1129.

[35] Stare, J., Henderson, R., and Pohar, M. (2005). An individual measure of relative survival. *Journal of the Royal Statistical Society — Series C*, 54(1):115–126.

[36] Stute, W. and Wang, J. L. (1994). The jackknife estimate of a Kaplan-Meier integral. *Biometrika*, 81(3):603–606.

[37] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1):20–22.

[38] Zadnik, V., Primic Žakelj, M., and Krajc, M. (2012). Cancer burden in Slovenia in comparison with the burden in other European countries. *Zdravniški vestnik - Slovenian Medical Journal*, 81(5):407–12.