Flexible Models for Competing Risks and Weighted Analyses of Composite Endpoints

Thesis

Flexible Models for Competing Risks and Weighted Analyses of Composite Endpoints

by

NGUYEN DUC ANH, BSc

A thesis submitted to the Open University UK in fulfilment of the requirements for the degree of

Doctor of Philosophy in the field of Mathematics and Statistics

conducted at

Oxford University Clinical Research Unit

Ho Chi Minh City, Vietnam

February 2015

Flexible Models for Competing Risks and Weighted Analyses of Composite Endpoints

by

NGUYEN DUC ANH, BSc

A thesis submitted to the Open University UK in fulfilment of the requirements for the degree of
Doctor of Philosophy in the field of Mathematics and Statistics

conducted at

Oxford University Clinical Research Unit
Ho Chi Minh City, Vietnam

February 2015

Abstract

In many clinical studies the occurrence of different types of disease events over time is of interest. For example, in cardiovascular studies, disease events such as death, stroke or myocardial infarction are of interest. As another example, in central nervous system infections such as cryptococcal meningitis, unfavourable events such as death or neurological events and favourable events such as coma or fungal clearance are relevant. In statistical terminology, competing risks refer to data where the time and type of the first disease event are analysed. Such data arise naturally if a nonfatal disease event is of interest but is precluded by death in a substantial proportion of subjects. Competing risks are the topic of the first four chapters of this thesis. An alternative approach used in many randomized controlled clinical trials is to combine different harmful events to a single composite endpoint. The analysis of trials with a composite endpoints is the topic of the fifth chapter. This thesis is organised as follows:

Chapters 1 and 2 are introductory chapters and provide an overview of statistical approaches to competing risks and semi-nonparametric (SNP) density estimation. Two concepts that form the basis for the work in Chapters 3 and 4 are introduced here: the cumulative incidence function (CIF) and SNP densities. For competing risks data, the CIF describes the absolute risk of different event types depending on time and is the most important quantity for data description, prognostic modelling, and medical decision making. SNP densities are densities that can be expressed as the product of a squared polynomial (of variable degree) and a base density which is chosen as the standard normal or the exponential density in this work.

Chapter 3 presents a novel approach to CIF-estimation. The underlying statistical model is specified via a mixture factorization of the joint distribution of the event type and time and the time to event distributions conditional on the event type are modelled using SNP densities. One key strength of the approach is that it can handle arbitrary censoring and truncation. A stepwise forward algorithm for model estimation and adaptive selection of SNP polynomial degrees is presented, implemented in the statistical software R, evaluated in a sequence of simulation studies, and applied to data sets from clinical trials in central nervous system infections. The simulations demonstrate that the SNP

approach frequently outperforms both parametric and nonparametric alternatives. They also support the use of "ad hoc" asymptotic inference to derive confidence intervals despite a lack of a formal mathematical verification for the relevant asymptotic properties.

Chapter 4 extends the work of Chapter 3 to regression modelling, i.e. the quantification of covariate effects on the CIF. A careful discussion of interpretational and identifiability issues which are intrinsic to models based on the mixture factorization is provided and the usage of the model is only recommended in settings with sufficient follow-up relative to the timing of the events. A simulation study demonstrates that the proposed approach is competitive compared to common statistical models for competing risks in terms of accuracy of parameter estimates and predictions. However, it also shows that "ad hoc" asymptotic inference is only valid if sample size is large. The chapter also provides a suggestion for model diagnostics of the proposed model, an area that has been somewhat neglected for competing risks data.

Chapter 5 discusses the analysis of composite endpoints. A common critique of traditional analyses of composite endpoints is that all disease events are equally weighted whereas their clinical relevance may differ substantially. This chapter addresses this by introducing a framework for the weighted analysis of composite endpoints that handles both binary and time-to-event data. To address the difficulty in selecting an exact set of weights, it proposes a method for constructing simultaneous confidence intervals and tests that protect the familywise type I error in the strong sense across families of weights which satisfy flexible inequality and order constraints based on the theory of $\bar{\chi}^2$-distributions. It is then demonstrated in several simulation scenarios as well as applications that the proposed method achieves the nominal simultaneous overall coverage rate with lower efficiency loss compared to the standard Scheffe's procedure.

Final remarks are given in Chapter 6 together with an outlook for potential future research directions.

Acknowledgement

To many a PhD is a journey; to me this journey is also a growing process, at the end of which one becomes not only a better researcher but also a better person. Like other journeys, one often does not walk alone but rather with some or (with some luck) many companions, for which I am not an exception. I would like to dedicate this very early page to attributing the successful completion of my PhD thesis to those who have, in various ways, walked with me during my journey.

First and foremost I would like to express my sincere and deepest gratitude to my supervisor Dr. Marcel Wolbers, without whom this thesis could not be in this finest state. Your great advice and priceless support both academically and non-academically have broadened my knowledge in the various aspects of both statistics and life.

I would also like to thank my co-supervisor Professor Paddy Farrington for his admirable wisdom through his helpful and constructive comments on my thesis, despite us having never met in person till the day these words were written.

I am also using this opportunity to thank Professor Jeremy Farrar for his thoughtful guidance as my former co-supervisor, and for making OUCRU-VN a paradise for researchers, of which I have also been a beneficiary.

My special thanks for the following people for their tremendous scientific contributions to my thesis. First, I am grateful to Professor Jeremy Day and Dr. Estee Torok for granting me access to their data sets, which have made beautiful illustrations in my thesis. Second, I would also like to send my honest appreciation to Professor Michael Patriksson for his generosity and kindness in replying to several of my technical questions.

I would also like to show my gratitude to Training Department and Training Committee of OUCRU-VN for making the decision to fund my PhD project. My special thanks to Dr. Mary Chambers and Ms. Le Thi Kim Yen for their tremendous help in many administrative as well as other aspects involving my PhD.

To my colleagues in the joint group of Mathematical Modelling and Biostatistics, especially Dr. Maciej Boni, Phung Khanh Lam, Stacy Todd, Ho Thi Nhan, Le Thanh Hoang Nhat, Dao Nguyen Vinh,

Ha Minh Lam, Nguyen Thi Duy Nhat, Tran Dang Nguyen and Tran Thi Thanh Phuong, I would like to thank you for being great friends and making me treasure my time at OUCRU.

Professor Abdel Babiker and Professor Robin Henderson graciously agreed to be my external examiners. Together with Professor Paddy Farrington and Professor Kevin McConway who chaired my thesis oral examination, they gave me a fair and yet memorably enjoyable viva.

Finally yet most importantly, to my family words are not enough to express how grateful and in debt I am to you, especially my dearest and devoted mother Ly Ngoc My and my beloved father Nguyen Duc Son. You have made the most sacrifices for me during this PhD and in my whole life. To end this, I would like to send the sweetest words to my little love Chau Thuy Trang, with whom I would love to share not just this journey but also the ones to come.

Nguyen Duc Anh

Oxford University Clinical Research Unit

Viet Nam, January 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction to competing risks

Survival analysis plays an important role across a large spectrum of research fields such as industrial reliability testing and clinical research. The outcome in survival analysis is the time from a time origin until the occurrence of the event of interest such as failure of a component in industrial settings or death in clinical studies.

In many settings, however, subjects can experience different types of events over time. In statistical terminology, competing risks refers to the situation where the time and the type of the first occurring event is of interest. For example, a competing risks endpoint for patients with cryptococcal meningitis initiating anti-fungal therapy is the time from initiation of therapy to fungal clearance (beneficial event) or to death prior to fungal clearance (competing harmful event). Work on competing risks started in the 18th century (Putter et al. (2007)) when Bernoulli investigated the possible consequences on mortality rates caused by elimination of smallpox. However, models specific to competing risks have only been developed since around the 1970s (Gail (1975), Prentice et al. (1978) and Putter et al. (2007)).

According to Koller et al. (2012), there are two important clinical settings where competing risks prevail, i.e. included subjects are susceptible to several different disease events. The first setting refers to studies in elderly or multimorbid patients with long-term risk exposure such as smoking, diabetes or hypertension. The second setting are severely ill subjects with short term risk exposures such as acute infections, cell depletion, or mechanical ventilation which are frequently seen among intensive care unit patients, transplant recipients, or subjects with severe tropical diseases.

Andersen & Keiding (2012) gave a concise discussion and critique of current statistical approaches to competing risks. In particular, they highlighted three principles that a competing risks method should adhere to in their opinion, namely:

1. Do not condition on the future

2. Do not regard individuals having already experienced an event as remaining "at risk" of the other competing events

3. Stick to the real world

I shall elaborate more on each of these principles when discussing some of the well-known competing risks approaches along with their strengths and limitations. Before that I introduce some basic competing risks concepts.

1.1 Basic competing risks: notation and concepts

Given a total number of n subjects, the observed competing risks data for one subject consist of the time to the first event T and the type of that event D, possibly subject to censoring and truncation. Of note, here and elsewhere, I drop supscripting with the index i for referring to a specific subject i for simplicity unless really needed. The event type D can take one of J distinct values $\{1, ..., J\}$, where J is an integer denoting the total number of event types. $J = 1$ refers to ordinary survival analysis and $J = 2$ occurs in many competing risks applications. In the literature, the different event types are referred to as competing risks, competing events, competing causes or failure causes. The use of the last term may lead to confusion because an event can be both beneficial and harmful. I will generally refer to the event type $D = 1$ as the "event of interest" and the other events as "competing events". Besides marginal modelling of (T, D), the association of (T, D) with a set of baseline covariates $\mathbf{Z}$, i.e. regression modelling, is important but I will drop $\mathbf{Z}$ from all formulas for convenience unless it is really needed. Some authors have also studied time-varying covariates in the presence of competing risks e.g. Beyersmann & Schumacher (2008). However, modelling time-varying covariates $\mathbf{Z}(t)$ in competing risks data is beyond the scope of this thesis. In the following subsections the main quantities relevant to modelling the joint distribution of (T, D) are defined.

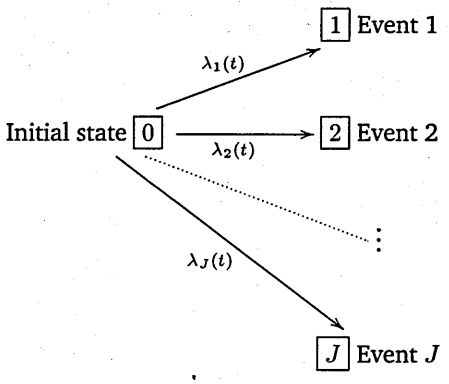
1.1.1 Basic competing risks entities

1. The cause-specific hazard (CSH) function or "force of transition" $\lambda_j(t)$ is defined as the instantaneous rate of having a specific event type j at time t conditional on being event-free up to time t (Aalen (1978))

$$\lambda_j(t) = \lim_{\Delta t \to 0^+} \frac{P\left(t \leq T < t + \Delta t,\ D = j \mid T \geq t\right)}{\Delta t} \tag{1.1.1}$$

Competing risks can be described by a multi-state model where each subject can move from the initial state (being event free) to one of the J event states with "transition intensities" (Beyersmann et al. (2012)) $\lambda_j(t),\ j = 1, ..., J$, see Figure 1.1.1. As is evident from Figure 1.1.1,

Figure 1.1.1: Multi-state model for competing risks.



competing risks methods do not consider further transitions between events, i.e. all event states are treated as absorbing. Of note, in some applications, transitions between events are possible (e.g. transition after 'fungal clearance' to 'death' in cryptococcal meningitis) but modelling them is outside the scope of competing risks and would require more general multi-state models (Beyersmann et al. (2012)).

2. The cumulative cause-specific hazard or "cumulative force of transition" $\Lambda_j(t)$ of event type j (Aalen (1978))

$$\Lambda_j(t) = \int_0^t \lambda_j(s)ds \tag{1.1.2}$$

3. The exponentiated negative cumulative cause-specific hazard

$$G_j(t) = \exp\left(-\Lambda_j(t)\right) \tag{1.1.3}$$

Of note, as raised in Putter et al. (2007), this is in general not the marginal survival function for the (latent) time to event type j in the absence of other event types. It only has this interpretation if the latent event times are independent of each other. This will be discussed further in Section 1.3

4. The overall survival function $S(t)$ and the total hazard function $\lambda(t)$ are given by

$$S(t) = P(T > t) = \exp\left(-\sum_{j=1}^{J}\Lambda_j(t)\right) = \prod_{j=1}^{J} G_j(t) \tag{1.1.4}$$

$$\lambda(t) = \sum_{j=1}^{J}\lambda_j(t) = \frac{-d\log S(t)}{dt} \tag{1.1.5}$$

These are the survival probability and the hazard of the time to the first event of any type (composite endpoint)

5. The cumulative incidence function (CIF) $CIF_j(t)$ of event type j measures the absolute risk (probability) of having an event of type j until time t.

$$\begin{aligned} CIF_j(t) &= P(T \leq t, D = j) \\ &= \int_0^t \lambda_j(s) S(s) ds \\ &= \int_0^t \lambda_j(s) \exp\left(-\int_0^s \lambda(v) dv\right) ds \\ &= \int_0^t \lambda_j(s) \exp\left(-\int_0^s \sum_{j=1}^{J} \lambda_j(v) dv\right) ds \end{aligned} \tag{1.1.6}$$

In the literature, the CIF is also named the probability of transition (Aalen (1978)), the cause-specific failure probability (Gaynor et al. (1993)), or the subdistribution function (Fine & Gray (1999)). Most competing risks approaches assume that all subjects will eventually have an event i.e. $\lim_{t\to\infty} \sum_{j=1}^{J} CIF_j(t) = 1$.

Amongst the quantities given above, the CSH and the CIF are the most prominent and form the target of most statistical models. The family of CSH functions $\lambda_j(t)$ or the family of CIFs, $CIF_j(t)$ for $j = 1, ..., J$, each uniquely determine the competing risks process or the joint distribution of (T, D) . However, as can be seen from (1.1.6) we see that the CIF of one event type j depends on the CSHs of all event types and not just event type j in a non-trivial way. This has implications on whether or not one should choose competing risks models based on the CSH or the CIF; especially for regression modelling as will be discussed in Section 1.5.

1.1.2 Censoring and Truncation

As in traditional survival analysis, competing risks data may also be subject to different types of censoring and truncation which prevent the observation of an event. Typical censoring and truncation schemes are right-censoring, left-truncation, and interval-censoring.

A subject is right-censored if it is known that the subject was still event free ("alive") at time t but the exact event time is unknown as the subject was not followed-up beyond time t. For the validity of traditional survival and competing risks models, it is essential that censoring is independent of the time-to-event process (see e.g. Kalbfleisch & Prentice (2002), page 194 for a precise definition) which essentially means that subjects censored at time t would have had the same future prognosis as subjects who remain under observation beyond time t. A simple form of independent right-censoring

occurs if a subject's censoring time C is stochastically independent of the competing risks process (T, D) and the observed data consists of the time to the event or censoring $\tilde{T} = \min(C, T)$, the censoring indicator $\delta = \mathbf{I}(T \leq C)$, and the censored event type $\tilde{D} = \delta \times D$, where $\mathbf{I}(p) = 1$ if the statement p is true and 0 otherwise.

Left-truncation or delayed entry occurs if some subjects only enter the study from time t onwards and some subjects are prevented from entering the study altogether because they have an (unobserved) event prior to entering. As for right-censoring, the left-truncation time must be independent of the time-to-event process for the validity of traditional survival and competing risks models.

An extension of right-censoring is interval-censoring which assumes that the subject's event is only known to have occurred in the interval $[L,\ R]$ but the exact time is unknown. If $R < \infty$, we shall always assume that the event type D is also observed. $R = \infty$ corresponds to right-censoring and in this situation the event type will also be unknown. Interval-censoring can be thought to occur because the subjects' status is only observed at certain time points characterized through a random observation process (M, V) where $M + 1$ denotes the number of observation times and $V = (V_0, ..., V_M)$ denotes the ordered times. The subject's event time might fall in between any two consecutive elements in V (Hudgens et al. (2014)) i.e $L = V_l$ and $R = V_{l+1}$ $(l = 0, ..., M - 1)$.

Formally there are two types of interval-censoring. The first type is "mixed case interval-censoring" which assumes the observation process to be independent of the time and the type of event i.e. $(M, V) \perp (T, D)$. For example, this situation occurs in clinical trials where the status is only assessed at pre-defined follow-up visits but an actual patient's visits may differ from the exact time point in a non-informative way. One special case of mixed case interval-censored data is "current status data" where each subject's event time is only known to occur either before or after a single random time point C. This means the interval is either $(0, C]$ or $[C, \infty)$ with $C < \infty$, where the event type can only be known in the former case while the latter corresponds to right-censoring. Such data usually arise from cross-sectional competing risks studies (Maathuis (2006)).

The assumption made in mixed case interval-censoring is not always valid; for example when at least one event type is death, there will not be any more observation times after death. This can be covered by the second type of interval-censoring called "independent inspection process" (IIP) which only requires the observation time V_l to be independent of the competing risks process conditional on the history of the observed data up to V_{l-1}. This means $V_l \perp (T, D) \mid H_l$ where H_l is the history of the observed data up to time V_{l-1} and $H_1 = V_0 = 0$ (see Hudgens et al. (2014) for details).

1.2 Naive approaches to competing risks and their criticism

An early naive approach to analysing the time to an event of interest in the presence of competing risks and right-censoring is to treat subjects experiencing competing events as right-censored at that time point and to then apply standard survival methods such as the Kaplan-Meier estimator and the log-rank test. Many authors (Gaynor et al. (1993), Gooley et al. (1999), Kalbfleisch & Prentice (2002) and Putter et al. (2007)) have criticized this approach because the Kaplan-Meier estimator is not a valid estimator of the CIF in the presence of competing risks. It incorrectly treats patients experiencing competing events as if they could have experienced the event of interest as a first event but that observation was precluded by censoring. Thus, the Kaplan-Meier estimator overestimates the CIF and examples where the naive Kaplan-Meier estimators of two competing events add up to more than 1 due to risk overestimation have been presented by several authors (Putter et al. (2007)).

It can be shown that the Kaplan-Meier estimator in the presence of competing risks estimates $1-G_j$ from Equation (1.1.3) rather than CIF_j (Lin (1997)). G_j can be interpreted as the marginal survival probability in a virtual world where the competing events could be prevented and this has no effect on the event of interest. However as emphasized in Gail (1975) and Putter et al. (2007), this poses a hypothesis that cannot be statistically tested because of the non-identifiability of latent failure times from the observed data as established in Tsiatis (1975). Of note, withdrawing subjects experiencing competing events from the risk set (i.e. treating them as right-censored) is a valid approach to non- and semi-parametric estimation of the (cumulative) CSH and associated regression models. Indeed, the classical nonparametric estimate of the CIF is based on (discrete) nonparametric estimates of all CSHs and then combining them according to formula (1.1.6), see Section 1.4.1 for more details.

In terms of statistical inference, log-rank tests have been used in the competing risks setting to compare CIFs between two groups. As noted in Lin (1997), this approach is flawed in the presence of competing risks, because the log-rank test for the event of interest actually compares the corresponding CSH between the 2 groups and not the CIFs. As the relation between the CIFs and the CSHs is non-trivial (see the last equality in Equation (1.1.6)), a reduction in the CSH of one cause does not necessarily translate into a reduction in the corresponding CIF (Putter et al. (2007)).

According to Koller et al. (2012) another common problem of published clinical studies in populations susceptible to competing risks is that they frequently focus on the event of interest exclusively and either do not report the competing event at all or only report its frequency without further analysis.

1.3 Approaches based on latent failure times

One way to estimate quantities of one event type while properly accounting for competing events is to consider competing risks data as a realization from a multivariate failure (event) time model, where each subject is assumed to have a potential (latent) failure time for each event type but only the time and type of the first occurring event are observed. The goal of such methods is to estimate the joint and marginal distributions of the latent times to event. As formulated in Gail (1975) and Putter et al. (2007), the joint multivariate survival distribution of the latent times to J distinct events is given by

$$\bar{S}(t_1, ..., t_J) = P\left(\bar{T}_1 > t_1, ..., \bar{T}_J > t_j\right) \tag{1.3.1}$$

where $\bar{T}_j$ denotes the latent time to event of type j and we only observe $T = \min_{j\in\{1,...,J\}}\{\bar{T}_j\}$ and $D = \text{argmin}_{j\in\{1,...,J\}}\{\bar{T}_j\}$. It follows that $\bar{S}(t, ..., t) = P(T > t)$ which is indeed the total survival function in (1.1.4). Moreover according to Theorem 1 of Tsiatis (1975), it is always true that $\frac{dCIF_j(t)}{dt} = -\frac{\partial \bar{S}(t,...,t)}{\partial t_j}$; and additionally $\lambda_j(t) = -\frac{\partial \log \bar{S}(t,...,t)}{\partial t_j}$ (Andersen & Keiding (2012)). Both the total survival function and the CSHs are identifiable from the observed data. The marginal survival function of event type j (called "net" survival probability in Tsiatis (1975)) is

$$\bar{S}_j(t) = P\left(\bar{T}_j > t\right) = \bar{S}(0, ...0, t, ..., 0) \tag{1.3.2}$$

Importantly $\bar{S}_j$ is only equal to G_j in (1.1.3) if the latent event times $\{\bar{T}_j\}_{j=\overline{1,J}}$ are assumed to be independent of each other. Indeed, the basic assumption of latent failure time approaches is the marginal independence of failure causes for general estimation or independence conditionally on covariates for regression modelling, see Gail (1975). Without the independence assumption, neither the joint survival function nor the marginal distributions are identifiable from the observed data as shown by Gail (1975) and Tsiatis (1975). Unfortunately, testing the independence assumption itself based on the observed data is also beyond the realm of statistics because for any independent latent failure times, a model with dependent failure times can easily be created which produces identical observed competing risks data (Tsiatis (1975)). As a simple illustration, assume that the two latent times $\bar{T}_1$ and $\bar{T}_2$ are independent and consider the alternative latent failure times pair $\bar{T}_1$ and $\tilde{T}_2 = \bar{T}_2\mathbf{I}\left(\bar{T}_2 \leq \bar{T}_1\right) + \left(\bar{T}_1 + \epsilon\right)\mathbf{I}\left(\bar{T}_2 > \bar{T}_1\right)$, $\epsilon > 0$. Clearly in the latter setting, the two latent times are not independent of each other; and yet the two settings yield the same observed data. In clinical settings, the independence assumption is often unrealistic and by using latent variables, this approach violates the principle to "stick to the real world" of Andersen & Keiding (2012).

One clinical area where latent failure time models have been used is in settings where one is

interested in estimating how the survival of a population would change if one cause of death, e.g. death due to cancer, could be eliminated (Andersen & Keiding (2012) and Honoré & Lleras-Muney (2006)). Once one is willing to make the independence assumption, one can e.g. estimate G_j with the Kaplan-Meier method considering subjects having competing risks as censored and then interpret G_j as the marginal survival function for event type j. Of note, this does not estimate CIF_j because $1 - G_j \neq CIF_j$ regardless of the independence assumption. The independence assumption might be more realistic in clinical setting if only independence conditional on carefully chosen covariates is assumed and using regression modelling instead of marginal inference can lead to more realistic estimates. However, Andersen & Keiding (2012) caution that in the clinical context, removing one disease from a population may also affect a person's susceptibility to other diseases. Moreover, a critique from the side of causal inference is that the effect of eliminating a specific cause of death on other death causes depends on the exact intervention to achieve this (e.g. a surgical intervention might have a different impact on other causes of death than a behavioural intervention) (Höfler (2005)).

One method to avoid the independence assumption is to resort to parametric models for $\bar{S}$. However, as pointed out in Gail (1975), the choice of such models cannot easily be justified from the data. Another approach based on latent failure times that avoids the independence assumption is to estimate bounds for the marginal survival probabilities. Unfortunately, the resulting bounds are often wide and not practically useful (Honoré & Lleras-Muney (2006)). To circumvent this Honoré & Lleras-Muney (2006) proposed an approach using parametric assumption for each latent failure time to tighten the bounds. There are also references from econometrics (Heckman & Honoré (1989) and McCall (1996)) showing that if one includes covariates, identification can be improved.

This section discussed the main features of approaches based on latent failure times to competing risks. The use of such methods allows us to answer questions that cannot easily be addressed using other approaches to competing risks such as the effect of eliminating certain competing events. However, all latent failure time methods are based on assumptions that cannot be verified using statistical tools.

1.4 Estimation of the cumulative incidence function

For exploratory and descriptive analyses of competing risks data, estimation and graphical display of the CIFs, possibly stratified by categorical covariates, is frequent. A commonly used method for such purposes is the nonparametric estimator of the CIF.

1.4.1 Nonparametric estimation

Based on the third equality of Equation (1.1.6), one can estimate a CIF by first estimating the total survival function and the CSHs and then plug these estimates into 1.1.6. For right-censored data, these quantities can be estimated nonparametrically. As shown in Putter et al. (2007), one can use the Kaplan-Meier estimator $\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$ of the time to the composite event (i.e. occurrence of any event) for S and nonparametric maximum likelihood estimates (NPMLE) for λ_j: $\hat{\lambda}_j(t) = \frac{d_{ji}}{n_i}$ for $t = t_i$ and 0 otherwise. In these formulas, t_is are the observed event times (from any cause) in the data set, n_i is the size of the risk set at time t_i, d_i is the total number of subjects experiencing an event at time t_i and d_{ji} is the total number of subjects experiencing an event of type j at time t_i. The CIF is then estimated as: $\widehat{CIF}_j(t) = \sum_{i:t_i \leq t} \hat{\lambda}_j(t_i)\hat{S}(t_{i-1})$. The Kaplan-Meier of S can also be calculated from the estimates $\hat{\lambda}_j(t)$. Therefore this approach indirectly estimates the CIF via estimating all CSHs at the observed time points.

Statistical inference and large sample theories for the above nonparametric estimator of the CIF have been discussed by several authors. Gaynor et al. (1993) discuss how confidence bands for several competing risks entities can be estimated based on Taylor-series expansions and the delta method. Using counting process and martingale theory, Lin (1997) proved that the nonparametric estimator of the CIF is consistent, and that a properly normalized version of it converges in distribution to a zero-mean Gaussian process with a covariance function whose consistent estimator is provided. Additionally, Lin (1997) developed a resampling technique to approximate the distribution of this process and to construct simultaneous confidence bands for the cumulative incidence curve as well as tests for between-group comparisons of CIFs.

The nonparametric estimator of the CIF defined above is the most frequently used descriptive statistic for competing risks data and plays a similar role as the Kaplan-Meier estimator for survival data. It is readily available in most statistics software. Moreover, it does not suffer from issues related to nonidentifiability as functions of the cause-specific hazard are always estimable (Prentice et al. (1978)). In addition, compared to parametric methods, the use of nonparametric methods poses no concern about model misspecification. However as noted in Gaynor et al. (1993), nonparametric models are saturated. Therefore they tend to yield estimates with less efficiency compared to parametric models. Benichou & Gail (1990) conducted a simulation study showing that substantial efficiency gains by using parametric models are possible. However the same simulation indicates that nonparametric methods should be used when we have no prior knowledge about the shape of the true CIF. Another issue concerning nonparametric models is that they do not allow for extrapolation beyond the last observed time point. Finally, while the mentioned nonparametric estimator is applicable

under right-censoring and can easily be adapted to incorporate left-truncation, it cannot be directly applied to interval-censored data. Hudgens et al. (2001) proposed two nonparametric methods to estimate the CIFs for competing risks data subject to interval-censoring and truncation. However, even in standard survival analysis where there is only one type of event, the nonparametric estimator of the survival function has non-standard asymptotic properties and slower than the usual $\sqrt{n}$-convergence in certain settings (Kalbfleisch & Prentice (2002)). Recently developed asymptotic properties for several nonparametric models for current status data can be found in Li & Fine (2013), Groeneboom et al. (2008*b*) and Groeneboom et al. (2008*a*).

Parametric methods can also be used for CIF estimation. As many of the proposed parametric methods can also be used in the regression setting, they will be discussed later in Section 1.5.3.

1.4.2 Nonparametric comparison of cumulative incidence functions

In many clinical settings, competing risks data can be grouped by a categorical covariate such as the assigned treatment arms in a randomized trial. A natural question arising from this setting is whether and to what extent the absolute risk of a specific event type over time, i.e. the CIF, differs between groups. I only discuss the comparison of two groups here and, without loss of generality, I focus on the first event type. Let the CIF for group k ($k = 1, 2$) of the first competing risk be denoted by CIF_1^k. The null and alternative hypotheses are

$$\begin{aligned} &H_0 : CIF_1^1(\cdot) = CIF_1^2(\cdot) \text{ vs.} \\ &H_A : \left[CIF_1^1(\cdot) \geq CIF_1^2(\cdot) \text{ or } CIF_1^1(\cdot) \leq CIF_1^2\right] \text{ and } CIF_1^1(\cdot) \neq CIF_1^2(\cdot) \end{aligned} \tag{1.4.1}$$

where the targeted alternative hypothesis H_A assumes "stochastic ordering" of absolute risks.

The two most popular test statistics addressing the stochastic ordering alternative in (1.4.1) are the integrated weighted difference (IWD) and a variant of the Kolmogorov-Smirnov test for the absolute risks. The IWD is defined as

$$\int_0^\tau W(t) \left[\hat{CIF}_1^1(t) - \hat{CIF}_1^2(t)\right] dt \tag{1.4.2}$$

and the Kolmogorov-Smirnov type test takes the form

$$\sup_{t \in [0,\tau]} W(t) \left|\hat{CIF}_1^1(t) - \hat{CIF}_1^2(t)\right| \tag{1.4.3}$$

where $W(\cdot)$ is a positive weight function, and τ is a suitably chosen time point which is usually the minimum of the largest observation times in each group (Pepe & Fleming (1991)). Compared to IWD the Kolmogorov-Smirnov type test is more more sensitive to large differences over a short time

period than moderate differences during longer periods which could be more clinically desirable. In practice, both test statistics in (1.4.2) and (1.4.3) are often based on the nonparametric estimates of the relevant CIFs resulting in corresponding "nonparametric" statistics.

Works on using the IWD to compare two cumulative incidence functions based on their nonparametric estimates include Pepe (1991), Pepe & Mori (1993), Bajorunaite & Klein (2007) and Bajorunaite & Klein (2008), who also discussed approaches based on the Kolmogorov-Smirnov type test. Using counting process theory and martingale central limit theorems, it was proven that the corresponding IWD is asymptotically normal with zero mean under the null hypothesis. Various variance estimators for this distribution were proposed (Pepe (1991) and Bajorunaite & Klein (2007)). An alternative approach to compute p-values based on resampling was proposed by Bajorunaite & Klein (2007) who adapted earlier work by Lin (1997) for constructing confidence bands for a single CIF over a certain time period.

For the nonparametric IWD-based test one role of the weight function $W(\cdot)$ is to stabilise the test statistic. Additionally we can use the weight function to target a more specific alternative hypothesis such as stressing on earlier or later differences.

Even though the stochastic ordering alternative in (1.4.1) is a broad alternative, an even more general alternative is $CIF_1(\cdot) \neq CIF_2(\cdot)$ which includes the case of "crossing CIFs". A more appropriate test statistic sensitive to this general alternative is a variant of Cramer von Mises's test (Schumacher (1984) and Pepe & Fleming (1989)), which is

$$\int_0^\tau W(t) \left[\hat{CIF}_1^1(t) - \hat{CIF}_1^2(t)\right]^2 dt$$

However, the null distribution of this statistic has the form of a linear combination of infinitely many independent χ^2 random variables which leads to a rather complex calculation for the p-value (Schumacher (1984) and Pettitt & Stephens (1976)). Moreover, the stochastic ordering alternative could be more clinically relevant as it indicates a homogeneous effect over time.

1.5 Regression modelling of competing risks

In most applications, we are not only interested in estimating the CIFs but also informal estimation and testing of the effects of covariates on the competing risks outcome. Both the CSH function and the CIF are the targets of prominent parametric and semi-parametric regression modelling approaches to competing risks. As mentioned earlier, from the second equality in Equation (1.1.6), the two quantities are related and both the knowledge of all CSHs and the knowledge of all CIFs, respectively, completely characterizes the competing risks process. However, Equation (1.1.6) also shows that the CIF of the

event of interest does not only depend on the CSH of that event but also on the CSHs of all competing events. Thus, covariate effects on the CSHs of the event of interest cannot be directly interpreted on the CIF scale and vice versa. A case study with only one binary covariate demonstrating this interpretational limitation of CSH-based analyses was given in Beyersmann et al. (2007).

The choice between choosing a model for the CSHs or a model for the CIF should thus be carefully made and should depend on the research question (Koller et al. (2012)). Several publications have argued that the CIF (which directly models the absolute risk of events) should be the target for prognostic modelling and medical decision making (Gail & Pfeiffer (2005) and Wolbers et al. (2009)). For a "mechanical understanding" of the underlying competing risks process, the interpretation of covariate effects on all CSHs is often the most informative analysis (Beyersmann et al. (2007), Putter et al. (2007) and Koller et al. (2012)). Popular regression models for the CSH and the CIF are briefly discussed below.

1.5.1 The semiparametric cause-specific hazards model

One of the most commonly used CSH-based regression approaches is the semiparametric CSH model based on Cox's proportional hazards model (Cox (1972)). This models the CSH of event type j given covariate values $\mathbf{Z}$ as $\lambda_j(t \mid \mathbf{Z}) = \lambda_{j,0} \exp\left(\mathbf{Z}^T \beta_j\right)$, where $\lambda_{j,0}$ is the base line CSH for event type j and $\mathbf{Z}$ and β_j are the covariate vector and the corresponding regression coefficients, respectively. According to Holt (1978), β_j can be estimated by maximizing the partial likelihood as in Cox (1972), thus allowing for the use of usual asymptotic methods for inference. The partial likelihood is

$$L = \prod_{j=1}^{J} \prod_{i=1}^{N} \left[\frac{\exp\{\mathbf{Z}_i \beta_j\}}{\sum_{h \in R_i} \exp\{\mathbf{Z}_h \beta_j\}} \right]^{\mathbf{I}(D_i = j)} \tag{1.5.1}$$

where N is the total number of subjects in the datasets, J is the number of competing events, and R_i is the risk set at the event time t_i of subject i consisting of all subjects without censoring or an event of any type prior to that time. Of note, the partial likelihood (1.5.1) factorizes into partial likelihood contributions for each cause-specific hazard. Thus, if the different cause-specific hazards models do not share parameters, estimation can proceed with standard Cox survival software including a single event type only and censoring subjects with competing events. More generally, cause-specific hazards models with shared parameters for different cause-specific hazards can also be fitted using standard software for stratified Cox models using a data duplication method, i.e. generation of an extended dataset with $N \times J$ rows containing one data row for each subject and event type, respectively (Putter et al. (2007))

One limitation of the proportional cause-specific hazards model is the proportionality assumption

inherited from Cox's original work. Fortunately, this can be dealt with by considering time varying covariates.

An alternative to semiparametric CSH modelling are parametric CSH models as laid out in Benichou & Gail (1990). These models can gain more efficiency if the parametric form is "correctly" specified.

1.5.2 Fine and Gray models

Fine & Gray (1999) proposed a direct semiparametric model for the CIF. This model belongs to a class of transformation models with a $\log\{-\log\{1-u\}\}$ transform applied to the CIFs. The method is implemented based on the concept of the subdistributional hazard (SH) of cause j, $\tilde{\lambda}_j(t)$, defined as the hazard function for the improper random variable $T^* = I(D=j) \times T + \{1 - I(D=j)\} \times \infty$ which implies that $\tilde{\lambda}_j(t)$ is equal to $-d\log\left(1 - CIF_j(t)\right)/dt$. This hazard is then assumed to follow a proportional hazards specification: $\tilde{\lambda}_j\left\{t;\mathbf{Z}\right\} = \tilde{\lambda}_{j0}(t)\exp\left\{\mathbf{Z}^T\beta_j\right\}$, where $\mathbf{Z}$ and β_j are the covariates and regression coefficients respectively. Importantly, this model implies the following model for the CIF: $CIF_j(t;\mathbf{Z}) = 1 - \exp\left[-\int_0^t \tilde{\lambda}_{j0}(s)\exp\left\{\mathbf{Z}^T\beta_j\right\}ds\right]$ and one can thus directly interpret covariate effects on the cumulative incidence scale. Here it is worth emphasizing that it is still difficult to deduce quantitative effects of covariates on the CIFs. However, it is clear that a covariate associated with an increased SH of event type j for higher covariate values is also associated with an increase in CIF_j, thus there is a direct qualitative interpretation.

In the absence of censoring and truncation β_j can be estimated by maximizing the corresponding likelihood defined as (Fine & Gray (1999)):

$$\tilde{L}\left(\beta_j\right) = \prod_{i=1}^{n}\left[\frac{\exp\left\{\mathbf{Z}_i\beta_j\right\}}{\sum_{k\in\tilde{R}_i}\exp\left\{\mathbf{Z}_k\beta_j\right\}}\right]^{\mathbf{I}(D_i=j)} \tag{1.5.2}$$

where $\tilde{R}_i$ is the risk set at event time t_i. Unlike the risk set in Equation (1.5.1) which includes only event-free subjects, the Fine and Gray model uses an alternative risk set definition $\tilde{R}_i$ which keeps subjects experiencing events other than those of type j in the risk set indefinitely. Under right-censoring, a modification of the score function of Equation (1.5.2) based on inverse probability of censoring weighting can be used to estimate β_j.

The Fine and Gray model is closely related to Gray's K-sample test (Gray (1988)), a commonly used test statistic for comparing CIF of an event of interest between groups, which is usually regarded as the competing risks counterpart of the log-rank test. This test also considers the null and alternative hypotheses specified in (1.4.1). For the case of two groups and assume that the first event is of interest, the involved test statistic is $z = \int_0^\tau W(t)\left\{\tilde{\lambda}_1^1(t) - \tilde{\lambda}_1^0(t)\right\}dt$, where $\tilde{\lambda}_1^0$ is the estimate for the SH of the first competing risk under the null hypothesis based on data pooled from all groups, and $\tilde{\lambda}_1^1$

is the estimate for the SH of the first competing risk based on data from the first group. According to Theorem 1 of Gray (1988) τ and $W(\cdot)$ are chosen such that $n^{-1/2}z$ is asymptotically normal. In practice, these choices may also be context dependent. Unlike the test based on the IWD statistic defined in (1.4.2), Gray's K-sample test does not have the power for covering the whole stochastic ordering alternative in (1.4.1). This is because Fine and Gray's estimation of the subdistributional hazards assumes proportionality, while H_A in (1.4.1) includes some cases of crossing SHs.

Separate Fine and Gray regression models can in principle be applied to different event types to obtain the respective covariate-dependent CIF estimates. This is both a simplification (because the models for different event types can be run independently) and a limitation. The limitation is that it is generally mathematically impossible for the proportional subdistributional hazards model to hold simultaneously for all event types. Moreover, in studies of two competing events with a limited follow-up duration, it can easily occur that the Fine and Gray model indicates that a covariate is associated with a higher CIF for both the event of interest and the competing event. However, this is incompatible with the fact that for $t \to \infty$, the two CIFs must add up to 1 for all covariate values. Moreover, it could be that at some time points, the predicted CIFs for the two event types add up to values exceeding 1 for certain covariate values.

Furthermore, the risk set definition associated with the SHs violates principle 2 'Do not regard individuals having already experienced an event as remaining "at risk" of other' of Andersen & Keiding (2012) . Indeed, SHs are somewhat artificial constructs which should not be interpreted as realistic rates. Nonetheless, if the focus is on how covariates affect the CIFs then this use of the SHs is just a pragmatic way to achieve the goal.

A final limitation of the Fine and Gray model is that it cannot cope with arbitrary censoring or truncation. The original proposal assumed right-censoring only (Fine & Gray (1999)) but the framework has recently been extended to left-truncation (Geskus (2011)). However, to my knowledge, variations of the Fine and Gray model for interval-censoring have not yet been proposed.

1.5.3 Parametric and mixture factorization models

Several parametric approaches to competing risks have been proposed. Compared to nonparametric and semiparametric models, these models rely on more restrictive distributional assumptions. However, if the parametric model is a good approximation to reality, they might allow for more efficient estimation. In addition, the inclusion of arbitrary censoring and truncation patterns including interval-censoring is much more straightforward in the parametric setting. In principle, parametric models for the CSHs can be formulated but the current literature mostly discusses models for the CIF. One way

to model the CIF is through a mixture factorization (Larson & Dinse (1985))

$$CIF_j(t) = P(T \leq t, D = j) = P(T \leq t \mid D = j)P(D = j) \tag{1.5.3}$$

The marginal probabilities $P(D = j)$ are the probabilities of eventually having an event of type j and can be modelled as having a multinomial distribution. The conditional probability $P(T \leq t \mid D = j)$ is proper and can be modelled by any distribution for non-negative continuous random variables. For example, Larson & Dinse (1985) modelled the hazard of T given $D = j$ with a piecewise constant hazards model. An alternative parametric model that has been suggested is the 3-parameter Gamma distribution (Checkley et al. (2010)). In the regression context, standard parametric survival models can be used to characterize the distributions of $T \mid D = j$ conditional on covariates, and a multinomial regression model for modelling the marginal distribution of D. Alternatively, Ng & McLachlan (2003) proposed a mixture model where the hazards of $T \mid D = j$ follow proportional hazards models whose baseline hazards are nonparametrically specified as step functions.

In view of the total survival function in Equation (1.1.4), (1.5.3) means $1 - S(t) = \sum_{j=1}^{J} P(T \leq t \mid D = j)P(D = j)$. By specifying different models for different conditional distributions $P(T \leq t \mid D = j)$, we can use a mixture of models to estimate $S(t)$, hence the name "mixture factorization"

An extensive analysis of parametric mixture factorization models including the discussion of large-sample properties of maximum likelihood estimators, test statistics and model existence and uniqueness can be found in Maller & Zhou (2002). One criticism of the mixture factorization is that by conditioning on D, it conditions on the eventual failure cause which lies in the future which violates principle 3 of Andersen & Keiding (2012) to not condition on the future. To circumvent this, Nicolaie et al. (2010) proposed an alternative factorization:

$$P(T \leq t, D = j) = P(D = j \mid T \leq t)P(T \leq t) \tag{1.5.4}$$

A problem with this factorization is that it is complicated to model $D \mid T$. In Nicolaie et al. (2010), $D \mid T$ is specified by the relative hazard: $\pi_j(t) = \frac{\lambda_j(t)}{\sum_{k=1}^{J} \lambda_k(t)}$ which is modelled by a time-dependent multinomial logistic model: $\pi_j(t) = \frac{\exp(\beta_j^T \mathbf{B}(t))}{1+\sum_{k=1}^{J-1} \exp(\beta_k^T \mathbf{B}(t))}$, where β_j is a row vector of p parameters and $\mathbf{B}(t) = (B_1(t), ..., B_p(t))$ is a predefined set of p functions of time.

Instead of using a mixture factorization model, one can also directly model the CIF with a parametric function. However, as $CIF_j(\infty) = \lim_{t\to\infty} P(T \leq t, D = j) \neq 1$ in general, we must use an improper distribution to model this quantity. In Jeong & Fine (2007) a generalized Gompertz

distribution was used to model the CIF

$$CIF_j(t) = 1 - \exp\left[\beta\{1 - \exp(\alpha t)\}/\alpha\right] \tag{1.5.5}$$

CIF_j characterizes an improper distribution function if $\alpha < 0$ and $|\beta| < \infty$. According to Jeong & Fine (2007) such a direct model of the CIF requires fewer parameters than a mixture factorization modelling approach and is thus more efficient. Additionally, Jeong & Fine (2007) claimed that their direct approach has a better chance to capture plateau patterns of the CIF and therefore is suitable for cure type models. However, as the Fine and Gray model, this approach estimates CIFs of different event types separately and thus may violate the constraint that their sum should not exceed 1. Also unlike the mixture factorization model, identifiability of the model in Jeong & Fine (2007) has not been well established.

As for survival data, parametric competing risks model can be extended to handle interval-censoring and left-truncation. An example where a parametric mixture factorization model is used to analyse interval-censored competing risks data can be found in Lau et al. (2008) and Lau et al. (2011).

1.6 Contrasting different regression approaches to competing risks

Section (1.5) summarizes several competing risks models, for which the target of modelling is either the CSH or the CIF and the choice between them depends on the research questions. As discussed, CSH-based regression models allow direct interpretation of covariate effects on the CSHs but not the CIFs. Many of these approaches specify a Cox's proportional hazard model for each CSH. As a result, such models can easily be fit to data using techniques already developed by Cox. However, as for the original Cox's model, they cannot easily be extended to handle interval-censoring. On the other hand, parametric CSH models can easily cope with different types of censoring at the price of making distributional assumptions.

When interest is in covariate effects on the CIFs, one standard approach is to use the Fine and Gray model which assumes proportional SH and is directly linked to the CIF. Nevertheless, one needs to be cautious when interpreting the SHs and their ratios because, as discussed, they are based on a non-intuitive risk set definition which regards subjects experiencing other event type as still at risk of the event of interest.

In some competing risks analyses, both Cox proportional CSH models and Fine and Gray models are used (Grambauer et al. (2010)). In such occasions, the estimated hazards ratios or subdistributional hazards ratios refer to different quantities, respectively, and hence may differ substantially from each other (Latouche et al. (2007), Beyersmann & Schumacher (2007) and Grambauer et al.

(2010)). Initial studies on the relationship between the CSHs and the SHs can be found in Latouche et al. (2007), Beyersmann & Schumacher (2007), Beyersmann et al. (2009) and Grambauer et al. (2010) which showed that the subdistributional hazard function is related to the CSH according to the following formula:

$$\tilde{\lambda}_j(t) = \left(\frac{S(t)}{1 - CIF_j(t)}\right) \lambda_j(t) \tag{1.6.1}$$

As a consequence of Equation (1.6.1), proportionality cannot hold for both types of hazards simultaneously. There have been several attempts to study the validity of fitting the Fine and Gray model to data following a Cox proportional cause-specific hazards model. In such case, the Fine and Gray model estimates the "least false parameter" (LFP). In some cases (Grambauer et al. (2010)), the LFP can still yield SHs ratios that are quite close to the true CSHs ratios for a specific event type j. Empirically, this usually happens if the covariate affects only the CSH of the event type j but not of other event types.

An alternative approach for modelling the CIF is through a mixture factorization as described in Section 1.5.3. This method offers flexible ways to specify and test for covariate effects: on the entire competing risks process, only on the marginal probabilities $P(D = j)$ or only on the conditional part $P(T \leq t \mid D = j)$ (Checkley et al. (2010)). This cannot be achieved if one models directly the CIF using some improper distribution. Moreover, according to Lau et al. (2011), mixture factorization models allow us to estimate the CSHs and SHs without worrying about the proportionality assumption which cannot hold simultaneously for both hazard types. This can be done using the following equations

$$\lambda_j(t) = \frac{dCIF_j(t)/dt}{S(t)}, \quad \tilde{\lambda}_j(t) = \frac{dCIF_j(t)/dt}{1 - CIF_j(t)} \tag{1.6.2}$$

In addition, one can easily incorporate left-truncation and interval-censoring when all parts of the mixture factorization are parametrically specified.

Chapter 2

Semi-nonparametric densities and their application in survival analysis

As discussed in Subsection 1.4.1, parametric models for competing risks might gain efficiency compared to non- or semi-parametric models, and can more easily incorporate arbitrary censoring and truncation patterns such as interval censoring at the cost of posing more restrictive distributional assumptions. Thus a method that combines the advantages of both approaches is desired. In this thesis, I investigate one potential approach to achieve this based on the so-called smooth "semi-nonparametric" (SNP) density representation introduced by Gallant & Nychka (1987) in an econometric setting. This method belongs to a wider class of "sieve extremum estimation" methods. Therefore, I shall begin this chapter by briefly discussing the general ideas of sieve estimation while restricting and focusing the exposition on aspects that are relevant to this thesis. Then I shall discuss the approach based on SNP densities as a special case of sieve estimation and discuss its applications in survival analysis.

2.1 Sieve extremum estimation

2.1.1 Introduction

In statistics, we often have a sample of n random (possibly multi-dimensional) variables $Z_1, ..., Z_n$, which are independently and identically distributed (i.i.d) according to a distribution determined by a "true" parameter $\theta_0 \in \Theta$. The space Θ can be of infinite dimension, for example a function space. The parameter space Θ is usually a metric space with a metric d and I will assume that such a metric exists throughout the rest of this section. To estimate θ_0 from the sample we need an empirical (sample-dependent) criterion function $\hat{Q}_n(\theta) : \Theta \to \mathbb{R}$. Estimation of θ_0 could then in principle proceed by maximizing $\hat{Q}_n$ over Θ yielding the estimator: $\arg\max_{\theta\in\Theta} \hat{Q}_n(\theta)$. However, when the

dimension of Θ is not finite, it is difficult to compute such a maximizer (Chen (2007) and Bierens (2014)). Moreover, it may be ill-defined, inconsistent or converge to the true parameter only at a slow rate of convergence (Chen (2007)). In his book Grenander (1981) proposed a solution to this problem by introducing the method of sieve extremum estimation. The idea is to optimize the criterion function over a sequence of much less complex and often finite dimensional parameter spaces $\{\Theta_n\}_{n=1}^{\infty}$ called sieves of increasing complexity. The sieves $\{\Theta_n\}_{n=1}^{\infty}$ usually refer to a non-decreasing sequence($\Theta_1 \subseteq \Theta_2 \subseteq \cdots \Theta_n \subseteq \cdots \subseteq \Theta$) of suitably chosen finite dimensional subspaces which are dense in Θ i.e. $\overline{\bigcup_n \Theta_n} = \Theta$.

In many applications including the semi-nonparametric models in this thesis, $\theta = (\xi, h) \in \Xi \times \mathcal{H}$, where Ξ is a subset of a finite dimensional Euclidean space and $\mathcal{H}$ is a function space of infinite dimension. In this case one may be interested in both components ξ and h or just in ξ with h being considered as a nuisance parameter. Estimation of ξ and h can be done using two main approaches: two-step procedure or simultaneous estimation. In the former, we first estimate h by $\hat{h}$, and then subsequently estimate ξ using $\hat{h}$ in place of h in the criterion function. Estimation of h in this case can be done nonparametrically or based on a sieve with the corresponding sieves $\{\mathcal{H}_n\}_{n=1}^{\infty}$ in $\mathcal{H}$. In the simultaneous estimation approach, estimation of ξ and h is done simultaneously by maximizing the criterion function over the sieve space $\Theta_n = \Xi \times \mathcal{H}_n$, which is what I will implement for my models.

The literature on sieve estimation is substantial and a large number of criterion functions and sieve spaces have been proposed. Indeed, many known estimation methods can be placed under the framework of sieves. Some examples are: the histogram and penalized regression (Geman & Hwang (1982)) as well as generalized least squares and maximum likelihood estimation (Chen (2007)). The choice of a suitable criterion function and sieve space depends on many considerations including how easy it is to optimize the criterion function for a sieve and how good the large sample properties of the resulting sieve estimator are. Of most relevance to this thesis is the method of sieve maximum likelihood estimation (MLE), a branch of a broader class called sieve M-estimation (Chen (2007)). In this case, the empirical criterion function is simply a log-likelihood function, defined as

$$\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) \tag{2.1.1}$$

where $l(\theta, Z_i)$ is the log-likelihood contribution of the i^{th} observation for a given θ. For future use define $Q(\theta) = E[l(\theta, Z)]$ where the expectation is taken under the "true" distribution of Z which is determined by θ_0. We can think of Q as a "population" or deterministic criterion function. The sieve

MLE (extremum) estimate is then defined as

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta_n} \hat{Q}_n(\theta) \tag{2.1.2}$$

Of note, in practice it is often impossible to find an exact MLE extremum estimator due to computational reasons, but rather an approximate maximizer which has the property: $\hat{Q}_n\left(\hat{\theta}_n\right) \geq \sup_{\theta \Theta_n} \hat{Q}_n(\theta) - O_P(\eta_n)$ with $\eta_n \to 0$ as $n \to \infty$. However, this does not change the asymptotic properties of the estimator. Thus for simplicity I consider $\hat{\theta}_n$ as in (2.1.2).

2.1.2 Large sample properties of sieve MLE

The literature on asymptotic or large sample properties of sieve estimators is extensive and of considerable technical sophistication with frequent usage of advanced methods from functional analysis and empirical process theory. As this topic is not the focus of this thesis and an extensive and relatively recent overview article exists (Chen (2007)), I shall restrict this short overview to general results on sieve maximum likelihood estimation which were accessible to me. Of note, asymptotic results specifically developed for SNP density estimators, the sieve estimators mostly relevant to this thesis, will be deferred to Section 2.2. In the following, I use definitions and notation from the previous section i.e. our data is modelled by i.i.d. observations $Z_1, ..., Z_n$ which follow the same distribution as Z. It is further assumed that the support of Z is in an Euclidean space. Additionally, I use $\text{plim}_{n\to\infty} X_n = X$ to mean that $X_n \underset{n\to\infty}{\longrightarrow} X$ in probability and $a_n \asymp b_n$ denotes that there exist constants c_1 and c_2 such that $c_1 \leq \frac{a_n}{b_n} \leq c_2$ for all n.

Consistency of sieve MLE

Consistency of sieve MLE has been discussed by several authors, either specifically or as a special case of sieve M-estimation, and a variety of regularity conditions required to establish consistency have been provided. Early discussions can be found in Geman & Hwang (1982) and Gallant & Nychka (1987). An overview article is Chen (2007), whereas Bierens (2014) provides a recent consistency proof under low-level conditions. Many of these results are applications or generalizations of Wald's classical consistency theorem which is included in most general textbooks on asymptotic theory (see e.g. Theorem 5.14 of Van der Vaart (2000)).

Theorem 3.1 of Chen (2007) states that the sieve estimator defined in (2.1.2) is a consistent estimator for the true parameter θ_0 under the following conditions:

- Condition 1 (identification) $Q(\theta)$ is uniquely maximized on θ at $\theta_0 \in \Theta$, and $Q(\theta_0) > -\infty$.

- Condition 2 (sieve spaces) $\Theta_n \subseteq \Theta_{n+1} \subseteq \Theta$, $\forall n \geq 1$; and there exists a sequence $\pi_n\theta_0 \in \Theta_n$ such that $d(\theta_0, \pi_n\theta_0) \to 0$.
- Condition 3 (continuity) The criterion function, $Q(\theta)$, is upper semi-continuous with respect to the metric $d(\cdot,\cdot)$.
- Condition 4 (compact sieve space) The sieve spaces, Θ_n , are compact under the topology implied by the metric $d(\cdot,\cdot)$.
- Condition 5 (uniform convergence over sieves) $\text{plim}_{n\to\infty} \sup_{\theta\in\Theta_n} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = 0$.

For $\Theta_n = \Theta$, these conditions essentially reduce to standard parametric conditions for the consistency of MLE over a compact space. However, the main use of this theorem is for cases where Θ is not compact which is true for several semi-nonparametric models including the ones developed in this thesis. Conditions 1 to 4 are standard regularity conditions but condition 5 is often difficult to verify. For this reason, Chen also discussed an alternative and potentially easier to verify condition, see condition 3.5M in Chen (2007). Bierens (2014) also criticised condition 5. Specifically in some MLE settings $Q(\theta)$ can be $-\infty$ for some θ^*. If such θ^* belongs to Θ_{N^*} for a large enough N^* while $\hat{Q}_n(\theta)$ is finite for all n, $\sup_{\theta\in\Theta_n} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = \infty$, $\forall n \geq N^*$ leading to $\sup_{\theta\in\Theta_n} \left|\hat{Q}_n(\theta) - Q(\theta)\right| \to \infty$ almost everywhere (a.s.). To overcome this, Bieren constructed his own consistency results with a new set of conditions as a generalization of Wald's theorem for the specific case of sieve MLE, see Assumption 4.1 and Theorem 4.1 in Bierens (2014). Finally, Theorem 3.1 in Chen (2007) already incorporates non-i.i.d observations, and Theorem 4.1 in Bierens (2014) can also be extended to such situation.

Convergence rates of sieve MLE

Similar to the study of consistency, several authors have studied convergence rates of sieve M-estimators in general and sieve MLE in particular, see for example Van de Geer (1993), Wong & Shen (1995), Birgé & Massart (1998) and Chen (2007). For simplicity I restrict the discussion here to results from Chen (2007). A general result on convergence rates for "usual" M-estimator, i.e. not in the sieve context, is provided in Corollary 5.53 of Van der Vaart (2000). According to this, the convergence rate depends on the behaviour of the empirical criterion function $\hat{Q}_n$ which can be written as the sum of the deterministic map Q and a random fluctuation $\hat{Q}_n - Q$. Intuitively, $\hat{\theta}_n$ converges rapidly to θ_0 if the deterministic map changes quickly as θ moves away from θ_0, while the random fluctuation remains small. Setting specific bound constrains to guarantee these behaviours and using the theory of

empirical processes, Van der Vaart showed that under suitable conditions, we can obtain the expected asymptotic result: $n^{1/2}d\left(\hat{\theta}_n, \theta_0\right) = O_P(1)$.

For sieve estimation, the rate of convergence depends on two quantities: 1) on the sieve approximation error rate $d(\theta_0, \pi_n\theta_0)$ (as defined in the consistency theorem above) and 2) on the complexity of the sieve space as measured by the so-called metric entropy with bracketing δ_n (*Wong & Shen (1995)* and *Chen (2007)*) taking values in $(0, 1)$. According to (Chen (2007)), for sieve M-estimators and i.i.d. data, $d(\theta_0, \pi_n\theta_0) = O_P(\epsilon_n)$ holds with $\epsilon_n = \max\{\delta_n, d(\theta_0, \pi_n\theta_0)\}$ under suitable regularity conditions. As n increases the complexity δ_n of Θ_n increases, whereas $d(\theta_0, \pi_n\theta_0)$ decreases. Thus, for an optimal rate of convergence, one should choose the complexity of the sieve spaces such that $\delta_n \asymp d(\theta_0, \pi_n\theta_0)$

Asymptotic normality of sieve MLE

Unlike consistency, there has not been a rich literature on asymptotic normality for sieve M-estimators, particularly when $\theta = (\xi, h) \in \Xi \times \mathcal{H}$ and simultaneous estimation as described in Subsection 2.1.1 is used. Nonetheless, some advances in this topic (Shen (1997) and Chen & Shen (1998)) were summarized in Section 4.2.1 of Chen (2007). In particular, Theorem 4.3 of Chen (2007) gives conditions for $\sqrt{n}$-asymptotic normality of smooth functionals of sieve M-estimators. However, according to Bierens (2014) these are complex conditions, because they cover a wide class of sieve M-estimators. Alternatively, he proposed lower-level conditions for deriving $\sqrt{n}$-asymptotic normality of the finite (parametric) component ξ of sieve MLE, see Section 6 of Bierens (2014). However, one of the associated assumptions is that $\Theta = \Theta_n$ for some n which is rather restrictive. Moreover, in his concluding remarks, Bierens (2014) admitted that he failed to verify one of these asymptotic normality conditions for his example, the SNP Logit model, which is simpler than the models discussed in this thesis.

2.2 Semi-nonparametric (SNP) density estimation

This section discusses a special type of sieve estimation, the so-called semi-nonparametric (SNP) density estimates which play an essential part in my thesis. Of note, whereas some publications discuss multivariate estimation, I simplified all statements and notations to the univariate case which is directly related to my work.

I start with laying down some definitions and notations useful for later discussions. For $p \in [1, \infty)$ and an integer $m > 0$ the Sobolev norm of a real valued function f with respect to a non-negative weight function $\mu(x)$ is defined as

$$\|f\|_{m,p,\mu} = \left(\sum_{\lambda \leq m} \int_{\mathbb{R}} \left| \frac{d^\lambda}{dx^\lambda} f(x) \right|^p \mu(x) dx \right)^{1/p} \tag{2.2.1}$$

for $p < \infty$ and as

$$\|f\|_{m,\infty,\mu} = \max_{\lambda \leq m} \sup_{\mathbb{R}} \left| \frac{d^\lambda}{dx^\lambda} f(x) \right| \mu(x)$$

for $p = \infty$. For $m = 0$ and $\mu = 1$ two special cases are the L_1 and L_2 norms for $p = 1, 2$ respectively. The normed linear space of real valued functions f with $\|f\|_{m,p,\mu} < \infty$ is denoted by $W_{m,p,\mu}$. For $p = 2$, $W_{m,2,\mu}$ is a Hilbert space with the corresponding inner product

$$\langle f, g \rangle_{m,2,\mu} = \sum_{\lambda \leq m} \int_{\mathbb{R}} f^{(\lambda)}(z) g^{(\lambda)}(z) \mu(z) dz$$

For a general reference on Hilbert spaces and Sobolev norms I refer to Adams & Fournier (2003).

SNP density estimation was first introduced in Gallant & Nychka (1987) for econometric applications in which estimation of some unknown parameters ξ_0 from a Euclidean space Ξ_0 of main interest often also requires estimation of an unknown density function f_0 based on a sample of n observations. If f_0 is assumed to have a parametric form so that it can be parametrized by a finite dimensional parameter θ_0, the method of maximum likelihood estimation (MLE) can be used to estimate ξ_0 and f_0 by optimizing the corresponding criterion function $\hat{Q}_n$ over the parameter space of ξ_0 and θ_0. However such a method can potentially produce biased results when the parametric assumption regarding f_0 is in doubt. To circumvent this Gallant and Nychka proposed a novel SNP density estimator inspired by the method of truncated Hermite series approximation. More specifically, Gallant and Nychka assume that the unknown density f_0 lies in a class $\mathcal{H}$ of "smooth", m_0 differentiable, densities defined as (Gallant & Nychka (1987) and Fenton & Gallant (1996*b*)) :

$$\mathcal{H} = \left\{ f_0(z) = g^2(z) + \epsilon_0 h_0(z) : \ \|g\|_{m_0,2,\mu_0} < \mathcal{B}_0, \ \|h_0\|_{m_0,2,\mu_0} < \mathcal{B}_0, \ \epsilon_0 > 0 \right\}, \tag{2.2.2}$$

where $\mu_0(x) = (1+z^2)^{\delta_0}$ and δ_0 is fixed and greater than $1/2$. h_0 is a fixed strictly positive, probability density function with expectation zero which together with the small positive constant ϵ_0 forms a lower bound for the tails of f_0; while $\mathcal{B}_0$ bounds the tails of f_0 from above and imposes smoothness restriction on f_0. These bounds are necessary to ensure that quantities such as $\int_{-\infty}^{\infty} \log f(z) f_0(z) dz$ and $\log f(z)$ are greater than $-\infty$ for all f and f_0 belonging to $\mathcal{H}$. According to Gallant and Nychka $\mathcal{H}$ accommodates f_0 with reasonable deviations from normal tails. This includes densities whose tails are not fatter than a t-like tail with $f_0(z) \propto (1+z^2)^{-\delta_0-\eta}$ and thinner than normal tail with $h_0(z) \propto e^{-u^{2+\Delta}}$ for some $\eta > 0$ and $\Delta \in (1, \delta_0 - 1)$. $\mathcal{H}$ is also flexible enough to include distributions

with any kind of skewness, kurtosis or multi-modality except for violent oscillations, kinks or jumps. Because $\mathcal{H}$ is an infinite dimensional parameter space, direct estimation of any $f_0 \in \mathcal{H}$ by optimizing a criterion function over $\mathcal{H}$ is difficult as discussed in Section (2.1.1). Therefore Gallant and Nychka proposed to simplify estimation of f_0 by looking at a much less complex, finite dimensional sieve space of the form

$$\mathcal{H}_n^0 = \left\{ f(z,\mathbf{a}) = \left(\sum_{j=0}^{K_n} a_j z^j\right)^2 b_0^2(z) + \epsilon_0 h_0(z) : \left\| \left(\sum_{j=0}^{K_n} a_j z^j\right) b_0(z) \right\|_{m_0,2,\mu_0} < \mathcal{B}_0 \right\}, \tag{2.2.3}$$

where h_0, ϵ_0 are as defined above, $b_0(z)$ is a strictly positive density function having a moment generating function, e.g. the normal density, and $\mathbf{a} = (a_j)_{j=0}^{K_n}$ must satisfy $\int_R f(z,\mathbf{a})dz = 1$. The degree of the polynomial K_n determines the complexity of $\mathcal{H}_n^0$ and its choice is often data-driven as discussed later in this section. As (2.2.3) uses the same ϵ_0 and h_0 as (2.2.2), it inherits the same tail conditions which can help to reduce computational problems like computing $\log f(z,\mathbf{a})$ when $P_{K_n}^2(z) = \left(\sum_{j=0}^{K_n} a_j z^j\right)^2$ is close to 0 in applications. According to Lemma A.5 in Gallant & Nychka (1987), using integration by parts $\mathcal{V}_0 = \left\{ P_K(z)b_0(z);\ \|P_K(z)b_0(z)\|_{m_0,2,\mu_0} < \mathcal{B}_0,\ K = 0,1,2,... \right\}$ is complete and therefore dense in $W_{m_0,2,\mu_0}$. As a consequence, for every g from (2.2.2), there exists an increasing sequence of polynomials $P_K(z)$ such that $\lim_{K\to\infty} \|g(z) - P_K(z)b_0(z)\|_{m_0,2,\mu_0} = 0$. This and Lemmas A.1 - A.3 in Gallant & Nychka (1987) imply that

$$\lim_{K\to\infty} \left\|g^2(z) - P_K^2(z)b_0^2(z)\right\|_{m_0,\infty,\mu_0} = 0, \tag{2.2.4}$$

which suggests that optimizing the log-likelihood $\hat{Q}_n(\xi, f)$ over $\Xi \times \mathcal{H}_n^0$ may yield good estimates for ξ_0 and f_0 when the degree K_n of the polynomial in (2.2.3) is increased with sample size. In fact Theorem 0 in Gallant & Nychka (1987) states that the estimates are consistent estimates for ξ_0 and f_0 under regularity conditions as stated below.

Theorem 0 *of* Gallant & Nychka (1987)

Consider $\Xi, \mathcal{H}, \mathcal{H}_n^0$ as introduced so far and $\hat{Q}_n$ as any finite sample based criterion function such as the log-likelihood function; let $\left(\hat{\xi}_n, \hat{f}_n\right) = \text{argmax}_{\Xi\times\mathcal{H}_n} \hat{Q}_n(\xi, f)$ be estimates of $\xi_0 \in \Xi$ and $f_0 \in \mathcal{H}$ respectively and let there be norms $|\xi|$ on Ξ and $\|.\|_{m,\infty,\mu}$ on $\mathcal{H}$, where $\mu(z) = \left(1+z^2\right)^\delta$ with $m \in (1/2, m_0)$ and $\delta \in (1/2, \delta_0)$.

(a) Compactness: The closure of $\mathcal{H}$ with respect to $\|.\|_{m,\infty,\mu}$ is compact in the relative topology generated by $\|.\|_{m,\infty,\mu}$

(b) Denseness: $\bigcup_{n=1}^{\infty} \mathcal{H}_n^0$ is a dense subset of the closure of $\mathcal{H}$ with respect to $\|.\|_{m,\infty,\mu}$ and $\mathcal{H}_n^0 \subset \mathcal{H}_{n+1}^0$.

(c) Uniform convergence: There are points $(\xi_0, f_0) \in \Xi \times \mathcal{H}$ and there is a function $Q(\xi, f, \xi_0, f_0)$ that

is continuous in (ξ, f) *with respect to* $\|(\xi, f)\| = \left(|\xi|^2 + \|f\|^2_{m,\infty,\mu}\right)^{1/2}$ *such that*

$$\lim_{n\to\infty} \sup_{\Xi\times\bar{\mathcal{H}}} \left|\hat{Q}_n(\xi, f) - Q(\xi, f, \xi_0, f_0)\right| = 0$$

almost surely.

(d) Identification: Any point $(\xi, f) \in \Xi \times \bar{\mathcal{H}}$ *with*

$$Q(\xi, f, \xi_0, f_0) \leq Q(\xi_0, f_0, \xi_0, f_0)$$

must have $|\xi - \xi_0| = 0$ *and* $\|f - f_0\|_{m,\infty,\mu} = 0$.

If conditions (a) - (d) hold and $\lim_{n\to\infty} K_n = \infty$ *almost surely then*

$$\lim_{n\to\infty} \left|\hat{\xi}_n - \xi_0\right| = 0, \text{ almost surely,}$$
$$\lim_{n\to\infty} \left\|\hat{f}_n - f_0\right\|_{m,\infty,\mu} = 0, \text{ almost surely.}$$

As long as it is assumed that the true unknown density f_0 is in $\mathcal{H}$ and that $\mathcal{H}$ and the sieves $\mathcal{H}_n^0$ are defined in (2.2.2) and (2.2.3), respectively, Gallant and Nychka showed that conditions (a) and (b) hold in general, while conditions (c) and (d) must be verified for each specific problem. For instance in Gallant & Nychka (1987) this was done for two econometric applications. As $\|.\|_{m,\infty,\mu}$ is a strong norm, (Gallant & Nychka (1987) and Fenton & Gallant (1996*b*)) consistency of $\hat{f}_n$ under this norm means that derivatives, moments and many functionals of f_0 can also be consistently estimated.

Theorem 0 in Gallant & Nychka (1987) only requires that $\lim_{n\to\infty} K_n = \infty$, which includes both deterministic increments, e.g. $K_n = n^{1/5}$, and adaptive choice of K_n, which uses additional information from the observed data besides the sample size. However for practical and theoretical reasons, the former is strongly disfavoured (Eastwood & Gallant (1991)). In many applications, K_n is chosen based on an information criteria such as Akaike's information criterion (AIC), the Bayesian information criterion (BIC) or Hannan–Quinn's information criterion (HQC), see (Zhang & Davidian (2008)). This means that the model is fitted for increasing values of K_n until the employed information criterion is optimized. However, a qualitative study of the performance of SNP density estimation of various densities from the Marron-Wand test suite suggests that in such settings, even SNP density estimates based on AIC quite frequently chose K_n that were too low (Fenton & Gallant (1996*b*)). In view of this, other approaches for selection of K_n based on cross-validation and on an estimate of the integrated squared error, $\int \left(\hat{f}(z) - f_0(z)\right)^2 dz$ were suggested which were found to be more appropriate but can be computationally intensive in some applications (Coppejans & Gallant (2002)).

Once K_n is fixed, $\mathcal{H}_n$ is just a parametric class of densities and standard MLE can be used for estimation. Thus, the SNP method is a special case of sieve-MLE estimation. On the other hand even though each $\mathcal{H}_n$ is a parametric class, the complexity K_n is determined by the observed data and as discussed earlier, the class $\mathcal{H}$ includes most "reasonable" densities, i.e. model misspecification with respect to f_0 is less of an issue for the SNP method compared to parametric approaches. Moreover, varying K_n also provides an informal way to test deviations of the estimated densities from the leading parametric term $b_0^2(z)$. Therefore one can think of SNP as lying in-between parametric and nonparametric methods.

Even though the original sieve $\mathcal{H}_n^0$ was given as in (2.2.3), in many applications $b_0^2(z)$ is replaced by the standard normal density, denoted by $\varphi(z)$ (Fenton & Gallant (1996*b*), Coppejans & Gallant (2002), Zhang & Davidian (2008) and Doehler & Davidian (2008)). The corresponding new sieves are:

$$\mathcal{H}_n = \left\{ f(z, \mathbf{a}) = \dot{P}_{K_n}^2(z)\varphi(z) + \epsilon_0 h_0(z) \right\}, \tag{2.2.5}$$

where all bound conditions, constants and functions involving $\mathcal{H}$ and the new sieve are as defined earlier except for the use of $\varphi(z)$ in place of $b_0^2(z)$. This change in sieve definition affects only condition (b) (denseness) in Theorem 0 of Gallant & Nychka (1987). Nevertheless from the proof of Theorem 2 in Gallant & Nychka (1987), $\bigcup_{n=1}^{\infty} \mathcal{H}_n^0$ is still dense in $\bar{\mathcal{H}}$ as long as an analogue of (2.2.4) holds i.e. for each g from $\mathcal{H}$, there exists an increasing sequence of polynomials $P_K(z)$ such that

$$\lim_{K\to\infty} \left\| g^2(z) - P_K^2(z)\varphi(z) \right\|_{m_0,\infty,\mu_0} = 0, \tag{2.2.6}$$

which in turn only requires that $\mathcal{V} = \left\{ P_K(z)\sqrt{\varphi(z)};\ \left\| P_K(z)\sqrt{\varphi(z)} \right\|_{m_0,2,\mu_0} < \mathcal{B}_0,\ K = 0, 1, 2, ... \right\}$ is dense in $W_{m_0,2,\mu_0}$. This in fact holds as long as the base density $\varphi(z)$ has the following property: $\left(\int \sqrt{\varphi(u)} du \right)^{-1} \sqrt{\varphi(z)}$ being a strictly positive density with a moment generating function, which holds for example when $\varphi(z)$ has the form *constant* $\times \exp\left(-\frac{z^p}{q}\right)$ for $p \geq 1$ and $q > 0$, see Appendix (A.1).

We can also rewrite $\mathcal{H}_n$ in terms of Hermite polynomials, see (Fenton & Gallant (1996*b*) and Kim (2007))

$$\mathcal{H}_n = \left\{ f(z, \theta) = \left(\sum_{j=0}^{K_n} \theta_j \bar{H}_{e_j}(z) \right)^2 e^{-z^2/2} + \epsilon_0 h_0(z),\ \theta \in \Theta_n \right\}, \tag{2.2.7}$$

where $\Theta_n = \left\{ \theta = (\theta_0, ..., \theta_{K_n}) : \sum_{j=0}^{K_n} \theta_j^2 = 1 - \epsilon_0 \right\}$ and $\left\{ \bar{H}_{e_j}(z) \right\}_{j=0}^{\infty}$ are orthonormal, bounded Hermite polynomials and hence complete in $W_{0,2,\exp(-x^2/2)}$. For more details on Hermite polynomials, I refer to Section 3 in (Coppejans & Gallant (2002)). Expression (2.2.7) means each member of $\mathcal{H}$ can be

expressed as an infinite Hermite expansion and can be estimated by truncated Hermite expansions in (2.2.7). From a practical point of view the Hermite parametrization in (2.2.7) is also computationally convenient and stable (Fenton & Gallant (1996*b*)).

The standard normal density $\varphi(z)$ in (2.2.5) is also called the base density of the sieve. In principle, other parametric choices of the base density than the standard normal density are possible and have been applied in practical applications (Kim (2007), Zhang & Davidian (2008) and Doehler & Davidian (2008)). The only requirement on the base density is that (2.2.6) can still be attained. Ideally, the base density should be chosen to be "close" to the true underlying density as then a good approximation can be achieved with a low K_n. Additionally, the term $\epsilon_0 h_0(z)$ is often omitted in practical applications and this does not seem to cause any major issues; alternatively, one can set ϵ_0 to a very small number for which $\hat{Q}_n$ can still be computed without error and h_0 is then often chosen to be $\varphi(z)$ (Gallant & Nychka (1987), Gallant & Marie Davidian (2010), Coppejans & Gallant (2002), Doehler & Davidian (2008) and Zhang & Davidian (2008)).

Besides consistency , other asymptotic properties of SNP-based estimation methods have also been studied which I shall briefly review. In Fenton & Gallant (1996*a*), a slight modification of the SNP method was introduced with the use of the same sieve $\mathcal{H}_n$ in (2.2.5 or 2.2.7) but to approximate a different class of function:

$$\mathcal{H}_\infty = \left\{ f_0(z) = g^2(z)e^{-z^2/2} + \epsilon_0\varphi(z) : \; g \in W_{m_0,2,\exp(-z^2/2)}, \; \int_{\mathbb{R}} f_0(z)dz = 1 \right\}, \tag{2.2.8}$$

where $g(z)$ must additionally satisfy that for every $a_0 > 0$ and a_1, there exists $k_0, k_1 > 0$ such that $\int_{z^2 \geq a_0 B + a_1} g^2(z) \exp\left(-z^2/2\right) dz \leq k_0 \exp\left(-k_1\sqrt{B}\right)$. This is a more restricted tail condition than the condition for membership in $\mathcal{H}$ (Fenton & Gallant (1996*b*)). Fenton & Gallant (1996*b*) established results for the convergence rate of SNP estimators for members from $\mathcal{H}_\infty$ under the L_1 norm in the case of density estimation which also imply strong consistency under the same norm. Although the authors could not theoretically establish that the rate is asymptotically equivalent to the rate achieved by the kernel density estimator, Monte Carlo simulations by the same authors demonstrated that the performance of the SNP estimator is often qualitatively similar to the kernel estimator (Fenton & Gallant (1996*b*)). As stated by these authors, the L_1 norm is the natural norm for density estimation. Furthermore Coppejans & Gallant (2002) considered the same space $\mathcal{H}_\infty$ and proposed a method for selecting K_n based on the integrated squared error and cross-validation, and established the asymptotic validity of their approach with a convergence rate result of the SNP estimator under the Hellinger distance. In a different application, Kim (2007) derived the asymptotic distribution of a Kullback-Liebler type test statistic for comparing two densities on a compact support and established

a uniform convergence rate of a SNP estimator for a truncated density with compact support.

To my knowledge, no publication has yet formally demonstrated the asymptotic normality of SNP-based estimators under the exact setting of Gallant & Nychka (1987). However, simulation studies for several applications have demonstrated that standard errors and confidence intervals derived from parametric maximum likelihood inference, i.e. ignoring the adaptive choice of the complexity K_n and treating it as fixed, usually yield valid inferences (Gallant & Tauchen (1993), Gallant & Marie Davidian (2010), Zhang & Davidian (2010) and Zhang & Davidian (2008)). Furthermore, Eastwood & Gallant (1991) used a truncated Fourier series that is similar in spirit to the SNP method to estimate a periodic function which formed a component in an additive model. In this setting, the authors could prove that estimates of derivatives up to a certain order of the estimated function were asymptotically normal under both deterministic and adaptive choice of K_n, the length of the truncated Fourier series.

2.3 SNP densities in survival analysis

As mentioned in the previous section, SNP estimation was originally developed to solve econometric problems. However, more recently several SNP-based methods for biometrical applications have been proposed. For example Zhang & Davidian (2010) used SNP distributions in place of the standard normal for flexible modelling of the random effects in linear mixed effects models which are frequently used for analysing longitudinal data. More closely related to this thesis is the use of SNP method for analysing survival data proposed by Zhang & Davidian (2008) and Doehler & Davidian (2008). In these works a continuous, positive time-to-event random variable T_0 is modelled as

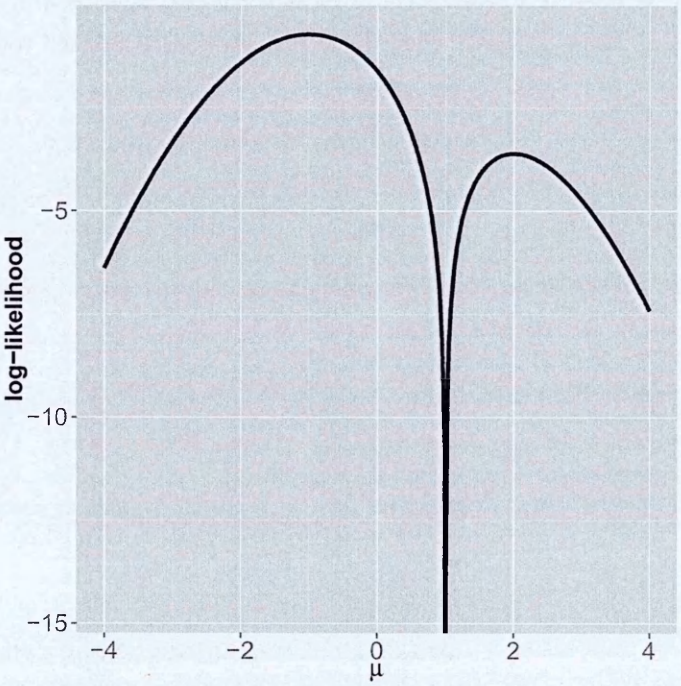
$$\log(T_0) = \mu + \sigma Z,\ \sigma > 0 \tag{2.3.1}$$

where Z is a random variable with support $(-\infty, \infty)$ and either Z or $\exp(Z)$ has a density f_0 in $\mathcal{H}$ defined in (2.2.2) which can be approximated by a density f_{K_n} coming from $\mathcal{H}_n$. Similar to many econometric applications the authors ignored the term $\epsilon_0 h_0(z)$ in the definition of $\mathcal{H}$ and demonstrated in their simulation studies that this raised no issues. Thus the SNP density estimator has the form

$$f_{K_n}(z) = P_{K_n}^2(z)\psi(z) = \left(\sum_{i=0}^{K_n} a_i z^i\right)^2 \psi(z) \tag{2.3.2}$$

where $\psi(z)$ is the base density. Zhang & Davidian (2008) and Doehler & Davidian (2008) recommended using either a standard normal base density when modelling the density of Z or an exponential base density if $\exp(Z)$ is modelled by a SNP distribution, respectively. For $K_n = 0$, this implies that T_0 has either a log-normal or a Weibull distribution. Therefore (2.3.2) provides a good approximation

Figure 2.3.1: Log-likelihood of μ: $\log\left\{\frac{1}{t}\left[\sqrt{2}+\sqrt{2}\left(\log t-\mu\right)\right]^2\frac{1}{\sqrt{2\pi}}\exp\left(-\left(\log t-\mu\right)^2/2\right)\right\}$ for a single observation with $t=1$.



for low-degree polynomials if the true density is relatively "close" to one of these two frequently used parametric distributions in survival analyses.

Although ignoring $\epsilon_0 h_0(z)$ did not bring any troubles in the simulation studies of Zhang & Davidian (2008) and Doehler & Davidian (2008), theoretically this can lead to a log-likelihood value of $-\infty$ which is numerically undesirable. A simple example is when we use model (2.3.1) with Z following the distribution $f_0(z)=\left(a_0+a_1z\right)^2\frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right)$. Then any combination of μ,σ and t that gives $\frac{\log t-\mu}{\sigma}=-\frac{a_0}{a_1}$ will cause the log-likelihood to be $-\infty$. Another potential issue involving the SNP log-likelihood is its multi-modality which was briefly mentioned in Zhang & Davidian (2008) but not strongly emphasized. However, it triggered these authors to consider multiple starting values for their model estimation. Multi-modality is a complication of most log-likelihoods based on SNP densities as illustrated in a simple example where $\log\left(T_0\right)=\mu+\sigma Z$ with $f_0(z)=\left(\sqrt{2}+\sqrt{2}z\right)^2\frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right)$, σ is fixed at 1 while μ is the only varying parameter. The corresponding likelihood $f_{T_0}\left(t\mid\mu\right)=\frac{1}{t}\left[\sqrt{2}+\sqrt{2}\left(\log t-\mu\right)\right]^2\frac{1}{\sqrt{2\pi}}\exp\left(-\left(\log t-\mu\right)^2/2\right)$ for a single observation $t=1$ is displayed in Figure 2.3.1. From this figure one can also see that the log-likelihood for μ is $-\infty$ when $\mu=1$.

In both Zhang & Davidian (2008) and Doehler & Davidian (2008), the SNP density f_{K_n} was parametrized using a set of spherical coordinates $\phi_{K_n}=(\phi_1,...,\phi_{K_n})$, with $\phi_i\in(-\pi/2,\pi/2]$, $i=$

$1, \ldots, K_n$. Web appendix of Zhang & Davidian (2008) gave details about these, which was also reproduced in Appendix A.2 of this thesis. One advantage of this parametrization is that it allows for a convenient way of setting initial values for optimization by choosing them from a regular "grid" over the parameter space. I shall also adopt this technique to my competing risks models with a slight modification and will discuss this in more details later. The authors chose the complexity of the SNP polynomial, K_n, based on an information criterion such as AIC, BIC or HQC and used MLE for fixed K_n.

Doehler & Davidian (2008) focused on flexible SNP-based estimation of the survival function and proposed test statistics for two-group comparisons of survival functions. Zhang & Davidian (2008) discussed the more general setting of regression models for survival data. Specifically, they used 2.3.1 as a model for the baseline survival or hazards function and additionally assumed that covariates affected this distribution according to either a proportional hazard model, an accelerated failure time model, or a proportional odds model. As SNP-based models are parametric for fixed K_n these authors could easily deal with different types of censoring and truncation without compromising model estimation. Moreover, their simulation results indicate that the resulting estimates frequently outperform both parametric and nonparametric alternatives. The simulation results from both publications further support the use of standard MLE inference which ignores the adaptive choice of K_n. While somewhat ad hoc, the simulation results show that this approach usually yields confidence intervals and significance levels which closely approximate the nominal values. For small sample sizes, they suggest the use of the bootstrap and show that this yields improved coverage of confidence intervals at the expense of increased computation time.

Even though the above works lack a strong theoretical justification for the use of SNP-based inference for survival analysis, the strong performance of SNP-based estimator compared to frequently used alternative survival models in simulation studies, the ability of the method to deal with arbitrary censoring and truncation patterns, and the successful use of SNP in earlier econometric applications provide strong support and motivation to extend these methods to the competing risks setting.

Chapter 3

SNP estimation of the cumulative incidence function

As outlined in the previous chapter semi-nonparametric (SNP) densities have been successfully applied in econometrics for more than two decades and have more recently been introduced to survival analysis. Motivated by the successful use of SNP methods in these areas which showed its advantages over traditional parametric, semi- and nonparametric methods and the growing importance of competing risks in medical applications, one aim of this thesis is to develop novel SNP competing risks models. The topic of this chapter is SNP estimation of the cumulative incidence function (CIF) whereas the next chapter covers regression modelling.

3.1 Model formulation

As in Chapter 1, competing risks data are denoted by (T, D) where T is the time to the first event and D is one of J discrete event types. The CIF for event type j is central to competing risks modelling and defined as $CIF_j(t) = P(T \leq t, D = j)$.

My SNP model for the CIF is based on a mixture factorization which characterizes the CIF as

$$P(T \leq t, D = j) = P(T \leq t \mid D = j)P(D = j) \tag{3.1.1}$$

This factorizes the CIF into a product of the marginal probability of the event type j and the probability of surviving up to time t conditional on eventually experiencing an event of type j. This factorization has been used by several authors (Larson & Dinse (1985) and Lau et al. (2011)) but has also been criticized because the conditional probability conditions on the future which is considered undesirable (Nicolaie et al. (2010) and Andersen & Keiding (2012)). I nevertheless use this factorization because

it is mathematically correct and provides a convenient factorization for statistical modelling. In particular, the goal of this chapter is to accurately estimate the CIF and I only use the factorization as a tool to achieve this.

Given $D = j$, $T \mid D = j$ is a "proper" random variable. Therefore the conditional probability $P(T \leq t \mid D = j)$ can be modelled with SNP densities. Specifically, I use the following "accelerated failure time" (AFT) model (Kalbfleisch & Prentice (2002) and Zhang & Davidian (2008)):

$$\log(T \mid D = j) = \log(T_{0j}) = \mu_j + \sigma_j Z_j \tag{3.1.2}$$

In the above equation Z_j is a random variable whose density is flexibly modelled by a SNP distribution. As in Zhang & Davidian (2008), I consider two different SNP models for Z_j: (1) a direct model of Z_j as a SNP density of degree K_j with a normal base density, and (2) an indirect model which models e^{Z_j} as a SNP density with an exponential base density with rate 1. Even though early research based on SNP densities exclusively used normal base densities, Zhang & Davidian (2008) argued that an exponential base density may be employed and the correspondingly fitted models can be compared based on suitable information criteria as discussed in the subsequent sections. For competing risks settings, different CIFs are estimated using different SNP densities which in general can have distinct base densities. This feature is also supported by the implementation of my competing risks models.

One of the benefits of SNP models is that for $K_j = 0$, the survival part of the models reduces to a standard parametric lognormal (for the normal base density) or Weibull (for the exponential base density) AFT model. Thus, the comparison of SNP models with $K_j = 0$ for all $j = 1, ..., J$ to more complex models with some $K_j > 0$ offers an informal test for the appropriateness of using more standard parametric models. Furthermore, when the degrees K_j of all SNP polynomials are fixed, the model is still parametric and standard maximum likelihood estimation and inference can be applied.

The marginal probabilities of my SNP models, namely $P(D = j)$ in (3.1.1), are assumed to follow a simple multinomial logistic model with intercept terms only:

$$P(D = j) = \frac{\exp(\gamma_j)}{1 + \sum_{k=1}^{J-1} \exp(\gamma_k)} \tag{3.1.3}$$

where γ_j $(j = 1, ..., J-1)$ are the parameters, and γ_J is set to 0 to ensure model uniqueness.

Importantly both model components (3.1.2) and (3.1.3) are flexible enough to incorporate covariates as discussed in Chapter 4 on regression modelling.

3.2 Likelihood construction

As discussed, my SNP model has two main components: the time-to-event distributions or (conditional) SNP components and the event type distribution or marginal component. For J different event types, J time-to-event distributions are required and will be parametrized by μ_j, σ_j, the chosen SNP base density, the degree of the SNP polynomial K_j and the coefficients of the polynomial. As in Zhang & Davidian (2008), I parametrize the polynomial coefficients by a set of spherical coordinates $\phi_{K_j} = \{\phi_{jk} \in (-\pi/2, \pi/2];\ k = 1, \ldots, K_j\}$ and review this parametrization in Appendix A.2.1. For the marginal component, $J-1$ additional parameters $\gamma_j (j = 1, ..., J-1)$ are required and the total parameter set for fixed base densities and a fixed set of SNP polynomial degrees $\mathbf{K} = \{K_1, ..., K_J\}$ is:

$$\gamma_j, j = 1, ..., J-1 \text{ and } \left\{\mu_j, \sigma_j, K_j, \phi_{K_j}\right\}_{j=1}^{J}$$

which I refer to as $\theta(\mathbf{K})$ for ease of discussion.

Simultaneous estimation of all parameters including the SNP polynomials degrees $\mathbf{K}$ is very difficult. However, for fixed SNP polynomial degrees, the method of maximum likelihood estimation (MLE), which is the most widely used method for statistical estimation and inference, can be used to estimate model parameters. Moreover, under the maximum likelihood framework, different censoring and truncation mechanisms can be included as for survival data (Zhang & Davidian (2008)). In this work, I consider competing risks models for data with right-censoring, interval-censoring and left-truncation, which are the most frequent types of censoring and truncation.

3.2.1 Likelihood contribution under right-censoring and left-truncation

Without left-truncation, a subject with an event of type j at time t contributes the following to the likelihood:

$$P(T \in [t, t+\delta t], D = j) = P(T \in [t, t+\delta t] \mid D = j) P(D = j)$$

where δt is an infinitesimally small positive number. Likewise, a right-censored case at time t contributes:

$$P(T > t) = 1 - \sum_{j=1}^{J} P(T \leq t, D = j) = 1 - \sum_{j=1}^{J} P(T \leq t \mid D = j) P(D = j)$$

In the presence of left-truncation at time lt, the respective likelihood contributions for an observed event or right-censoring, respectively, at time t with $t > lt$ are as follows:

$$\begin{aligned}
P(T \in [t, t+\delta t], D = j \mid T > lt) =& \frac{P(T \in [t, t+\delta t], D = j, T > lt)}{P(T > lt)} \\
=& \frac{P(T \in [t, t+\delta t], D = j)}{P(T > lt)} \\
=& P(T \in [t, t+\delta t] \mid D = j) P(D = j) \times \\
& \left\{ 1 - \sum_{k=1}^{J} P(T \leq lt \mid D = k) P(D = k) \right\}^{-1}
\end{aligned}$$

$$\begin{aligned}
P(T > t \mid T > lt) =& \frac{P(T > t, T > lt)}{P(T > lt)} \\
=& \frac{P(T > t)}{P(T > lt)} \\
=& \left\{ 1 - \sum_{j=1}^{J} P(T \leq t \mid D = j) P(D = j) \right\} \times \\
& \left\{ 1 - \sum_{k=1}^{J} P(T \leq lt \mid D = k) P(D = k) \right\}^{-1}
\end{aligned}$$

Of note, the likelihood expressions above ignored components for the censoring or left-truncation distributions which I assume to be independent of the distribution of the competing risks process. Hence, these components can be regarded as unknown constants which do not affect maximum likelihood estimation (MLE).

3.2.2 Likelihood contribution under interval-censoring and left-truncation

Under interval-censoring, the likelihood contribution of a subject with event type j occurring during the interval $[l, r]$ with $r < \infty$ is:

$$P(T \leq r, D = j) - P(T \leq l, D = j)$$

Under right-censoring i.e. $r = \infty$, and implicitly assuming that the event type is also unknown, the likelihood contribution is as before:

$$1 - \sum_{k=1}^{J} P(T \leq l, D = k)$$

Using the same rationale as before together with considering left-truncation the above formulas respectively equal:

$$\{[P(T \leq r \mid D = j) - P(T \leq l \mid D = j)]\, P(D = j)\} \times \left\{1 - \sum_{k=1}^{J} P(T \leq lt \mid D = k) P(D = k)\right\}^{-1}$$

and

$$\left\{1 - \sum_{k=1}^{J} P(T \leq l \mid D = k) P(D = k)\right\} \times \left\{1 - \sum_{k=1}^{J} P(T \leq lt \mid D = k) P(D = k)\right\}^{-1}$$

where the truncation time lt is smaller than l.

3.2.3 Log-likelihood calculations for fixed SNP polynomial degrees

Under the usual assumption of i.i.d observations, the log-likelihood of the entire sample equals the sum of the logarithm of all individual likelihood contributions as described above. Specifically for terms involving δt, I use the respective SNP densities. For terms having the form of $P(T \leq t \mid D = j)$ I evaluate the corresponding SNP survival functions. My implementation of the SNP density follows Zhang & Davidian (2008). This means the SNP density is as defined in Equation (2.3.2) which in turn follows directly Equation (2.2.5) with the tail $\epsilon_0 h_0(\cdot)$ omitted. In this formulation, both the standard normal and standard exponential are possible base densities.

As in Zhang & Davidian (2008) and Doehler & Davidian (2008), I did not enforce a bound on the Sobolev norm (Equation (2.2.1)) of $P_K(z)\psi(z)$, where $P_K(z)$ and $\psi(z)$ are the SNP polynomial and SNP base density, respectively. This could in principle be done by a priori fixing values of quantities determining the Sobolev norm, namely m_0 and δ_0, as discussed in the definition of the $\mathcal{H}$ class and its sieves in Section 2.2 but would require defining sensible default choices for these quantities and complicate the implementation. The only condition on the SNP densities that I enforced is that they must integrate to one, which is guaranteed by the spherical parametrization of the SNP polynomials.

Details of the spherical parametrization of the SNP polynomial as well as the calculations for the SNP densities and the corresponding survival functions are discussed in the web-based Appendix of Zhang & Davidian (2008) but can also be found in Appendix A.2 of this thesis.

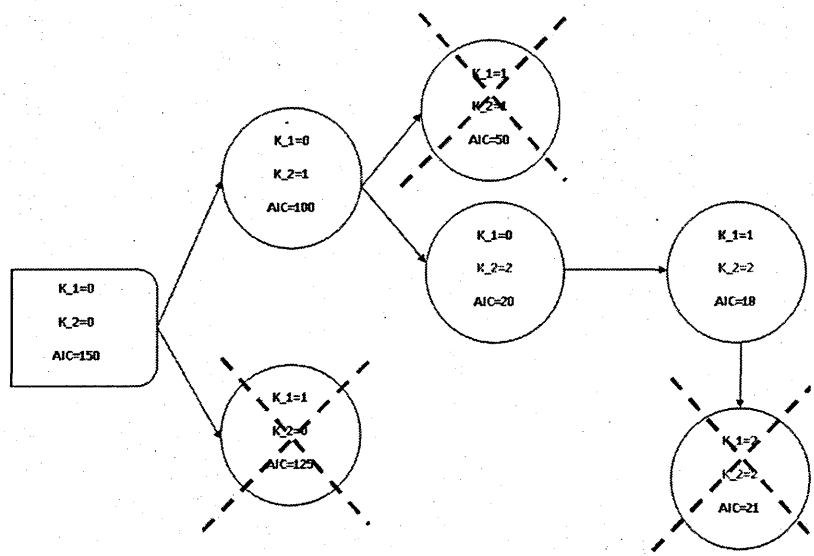
3.3 Estimation procedure

In the preceding section I have discussed the construction of the log-likelihood for a fixed set of polynomial degrees $\mathbf{K}$. In this case the other parameters of the model $\theta(\mathbf{K})$ can be estimated by MLE. However, including the polynomial degrees as unknown parameters into the log-likelihood is problematic because this would require optimization over an infinite-dimensional parameter space and

mandate optimization of the other parameters for all possible polynomial degrees which is infeasible. Moreover, as models with polynomial degrees are nested, models with higher degrees automatically tend to lead to better log-likelihoods. Hence, as discussed in Chapter 2, a more appropriate estimation technique is sieve extremum estimation with adaptive choice of the polynomial degrees according to an information criterion.

To reduce the computational complexity of the algorithm for multiple event types, I decided to implement this with a greedy step-wise forward algorithm and a fixed maximum complexity K_{max} for all polynomial degree $K_j (j = 1, \ldots, J)$. In my own implementation, I set $K_{max} = 2$ as this is already flexible enough to capture a wide range of distributions, but the algorithm is designed in a way that K_{max} can in principle easily be increased (at the expense of a more complex optimization problem). In the literature, values of K_{max} from 2 to 4 have been suggested based on simulations and practical applications (Zhang & Davidian (2008), Doehler & Davidian (2008) and Fenton & Gallant (1996*b*)). A schematic of the step-wise forward algorithm for 2 competing risks ($J = 2$), $K_{max} = 2$, and AIC as the information criterion is given in Figure 3.3.1. In this example, the final model would chose $K_1 = 1$ and $K_2 = 2$ as only one forward step is allowed at this point which does not further improve AIC.

Figure 3.3.1: Adaptive greedy step-wise forward selection of polynomial degrees K_1 and K_2 using AIC.



A fitted model with polynomial degrees $\mathbf{K} = (K_1, ..., K_J)$ is preferred to the other models based on an information criterion of the form: $2 \times \{-l_{\{K_1,...,K_J\}}(\theta(\mathbf{K})) + qc\}$; where $l_{\{K_1,...,K_J\}}(\theta(\mathbf{K}))$ is the log-likelihood evaluated at the MLE for fixed polynomial degrees and q is the total number of parameters in the model i.e. the dimension of $\theta(\mathbf{K})$. Three different information criteria were considered: AIC (Akaike's information criterion), BICn (Bayesian information criterion) and HQCn (Hannan-Quinn's

criterion) which set $c = 1, c = \log(n)/2$, and $c = \log\{\log(n)\}$, respectively. Moreover, I also considered replacing n, the sample size, by d, the total number of events of any type, in the definition of BIC and HQC and denote the resulting information criteria by BICd and HQCd, respectively (Volinsky & Raftery (2000)).

The greedy step-wise forward strategy for selecting the optimal polynomial degrees may miss the best possible information criterion. For example in Figure 3.3.1 although the model $(K_1 = 1, K_2 = 0)$ has a worse AIC, its "child node" $(K_1 = 2, K_2 = 0)$ may still have a better AIC than any of the considered nodes and the algorithm would have missed this. However, an exhaustive search through all models would require fitting $(K_{\max} + 1)^J$ models compared to at most $1 + J^2 (K_{max} - 1) + J(J+1)/2$ for the greedy algorithm. For example, for $K_{max} = J = 3$, an exhaustive search always requires 64 model fits while the greedy algorithm requires only 25 fits in the worst case.

For fixed polynomial degrees, the log-likelihood function is still rather complex and may have multiple extrema. Hence, good starting values for the numerical optimization algorithm are crucial. The next two subsections discuss how to initialize the algorithm for the parametric mixture factorization model corresponding to $K_1 = \ldots = K_J = 0$ and how to update parameter estimates from the previous step to obtain starting values for the new step with an increased polynomial degree. All algorithms and simulations in this chapter as well as in my entire PhD project were implemented in the statistical software R version 3.0.0 (2013-04-03) (R Core Team (2013)).

3.3.1 Starting values for the parametric mixture factorization model

In the parametric setting, the time-to-event for a specific event type is specified as: $\log(T_j) = \mu_j + \sigma_j Z_j$ leading to a log-normal model for T_j if Z_j has a standard normal base density and to a Weibull model if $\exp(Z_j)$ has a standard exponential distribution. Hence, initial values for μ_j and σ_j were obtained with standard software for fitting parametric AFT as outlined below.

Specifically, I used the R function `survreg` from package `survival` (Therneau & Grambsch (2000)), which supports survival data from both the lognormal and the Weibull distribution. All observations with an observed event type j and all censored observations were included in the AFT. Since the event type of a right-censored observation is unknown, they were not included "fully" (i.e. with weight 1) into the AFT log-likelihood. Rather, a right-censored observation at time t received a weight corresponding to its crude estimate of the probability of ultimately experiencing event type j given by $P(D = j \mid T > t)$. Note that $P(D = j|T > t) = P(D = j, T > t)/(P(D = 1, T > t) + \ldots + P(D = J, T > t))$ and $P(D = j, T > t)$, $j = 1, \ldots, J$ can be estimated as follows. First, the R function `cuminc` from package `cmprsk` (Gray (2013)) is used to get the nonparametric CIF estimates, denoted by $\hat{P}_{NP}(T \leq t, D = j)$. Second, the crude estimates $(\hat{P}_j)$ for the marginal event probabilities $P(D = j)$

are derived by normalizing the values of the CIF estimates at the last observed time point to 1. Then $P(D = j, T > t)$ is estimated by $\hat{P}_j - \hat{P}_{NP}(T \leq t, D = j)$.

For interval-censored observations, the event type is known, and they are included as interval-censored with weight 1 in the AFT. However, as the standard nonparametric estimator of the CIF does not allow for interval-censoring, interval-censored observations were assumed to experience the event at the mid-point of the interval for the sake of this estimate.

As suggested by Zhang & Davidian (2008) for survival analysis, in addition to using the starting values $(\mu_j, \sigma_j)_{j=1,\ldots,J}$ from the survival AFT, I also used $(\mu_j \pm \sigma_j/2, \sigma_j)_{j=1,\ldots,J}$ as additional sets of starting values.

To get initial values for $\gamma_j, j = 1, ..., J-1$, I optimized the full log-likelihood with respect to these parameters while fixing $(\mu_j, \sigma_j)_{j=1,\ldots,J}$ at each set of starting values mentioned above. This sub-optimization itself requires initial values for $\gamma_j, j = 1, ..., J-1$, which were taken to be $\hat{P}_j$ as described above. Then $\hat{P}_j$, $j = 1, \ldots, J$ were transformed to starting values for γ_j according to

$$\hat{\gamma}_j = \log\left(\frac{\hat{P}_j}{\hat{P}_J}\right), j = \overline{1, J-1} \tag{3.3.1}$$

For the sake of determining starting values, left-truncation is currently ignored. In principle, it would be possible to extend the heuristic techniques outlined above to this setting but current R implementations for parametric AFT and nonparametric CIF estimation do not support left-truncation which would complicate implementation.

3.3.2 Starting values for the intermediate step

At an intermediate step which increases $(K_1, \ldots, K_j, \ldots, K_J)$ to $(K_1, \ldots, K_j + 1, \ldots, K_J)$, initial values for $\gamma_j, j = 1, \ldots, J-1$ and all parameters related to conditional survival distributions for event types other than j will be set to the corresponding MLE values from the previous step.

For getting starting values for survival parameters related to event type j, I follow Zhang & Davidian (2008) who suggested to use a grid of starting values for the new spherical coordinates ϕ_{K_j+1} characterizing the SNP polynomial. Specifically, I use the grid $\{-1.5, -1.3, ..., 1.3, 1.5\}$ for $K_j + 1 = 1$ and $\{-1.5, -0.5, 0.5, 1.5\} \times \{-1.5, -0.5, 0.5, 1.5\}$ for $K_j + 1 = 2$. A relatively large number of starting values is chosen because the log-likelihood is expected to be multi-modal with respect to the spherical coordinates.

Besides the "default" grids for $K_j + 1 = 1$ and 2, in general for $K_j + 1 = m$ my implementation allows for a grid of the form $S_1 \times ... \times S_m$ whose $S_1, ..., S_m$ are sequences of equally spaced points in $[-1.5, 1.5]$ whose lengths must be predetermined.

Given initial values for ϕ_{K_j+1}, corresponding initial values for μ_j and σ_j are obtained as follows: First I calculate $\mathrm{E}(\log T \mid D = j)$ and $\mathrm{Var}(\log T \mid D = j)$ based on the MLE estimates from the previous step using the relation:

$$\begin{aligned} \mathrm{E}(\log T \mid D = j) &= \mu_j + \sigma_j \mathrm{E}(Z_j); \\ \mathrm{Var}(\log T \mid D = j) &= \sigma_j^2 \mathrm{Var}(Z_j) \end{aligned} \tag{3.3.2}$$

For the calculation of the moments of SNP distributions in (3.3.2), I refer to Appendix A.3. Second, given these values and the chosen new spherical coordinates, I solved (3.3.2) for μ_j and σ_j. In short, I update μ_j, σ_j to correspond to the new spherical coordinates ϕ_{K_j+1} such that the first two moments of $\log(T \mid D = j)$ remain unchanged.

3.3.3 Optimization

As mentioned above, for each step of the estimation procedure, numerical optimization of the log-likelihood is performed from a set of different starting values. As this is computer intensive, my implementation in the statistical software R uses parallelization as supported by the R package `parallel` (R Core Team (2013)).

For a given set of initial values I optimize the likelihood using function `maxLik` in the R package `maxLik` (Arne Henningsen & Toomet (2011)) which allows for different optimization approaches. After some experimentation, I chose the quasi-Newton method which calculates approximations to the Hessian using the "Broyden-Fletcher-Goldfarb-Shanno" (BFGS) updating formula. To improve numerical accuracy and speed up performance time, the gradient of the log-likelihood function was analytically implemented. Moreover, a direct calculation of the marginal probability in (3.1.3) is prone to computational instability even when the logarithm is calculated first because $\log P(D = j) = \gamma_j - \log\left(1 + \sum_{k=1}^{J-1} \exp(\gamma_k)\right)$ still has the terms $\exp(\gamma_k)$ which can potentially grow really fast. To overcome this, I instead calculate $\log P(D = j) = \gamma_j - A_{\max} - \log\left(\exp(-A_{\max}) + \sum_{k=1}^{J-1} \exp(\gamma_k - A_{\max})\right)$ with $A_{\max} = \max\{\gamma_1, \ldots, \gamma_{J-1}, 0\}$.

Amongst the fitted models for all starting values, I selected all obtained estimates for which the optimization algorithm suggested successful convergence and chose the parameters corresponding to the highest log-likelihood as the final estimate given the chosen degree of the SNP polynomials. In addition, for statistical inference, it is required that the best fit has a negative definite Hessian matrix. If this was not the case, I searched whether any fit from a different starting value had well-behaved Hessian and its corresponding log-likelihood was at most 1% less than the best fit. If this was the case, I used this alternative fit instead. If no solution with a well-behaved Hessian could be found, the

algorithm returns the best fits from the previous step with a warning message. Of note, the simulation study discussed in Section 3.6 avoided this check of the Hessian matrix to speed up computation. However, when final parameter estimates were examined, it turned out that degenerate Hessians were rare and only occurred in 3/2000 (0.15%) simulations with right-censored data and at most in 1% for a single scenario.

Besides the premature stopping of the algorithm due to a degenerate Hessian, the algorithm successfully terminates if the chosen information cannot be further improved or if all possible future steps would involve SNP polynomials with degrees exceeding K_{max}.

3.4 Ad hoc statistical inference for CIF estimates

As discussed in Sections 2.2 and 2.3, rigorous theoretical justification for asymptotic normality of SNP estimates is lacking in general. However, empirical evidence from several studies on density estimation and survival analysis suggest that the use of standard MLE-based inference for SNP estimates and associated functionals based on the final fit and ignoring the adaptive choice of the degrees of the SNP polynomials yields acceptable performance (Fenton & Gallant (1996*b*), Zhang & Davidian (2008) and Doehler & Davidian (2008)).

Accordingly I base the calculation of the standard errors and confidence intervals for quantities derived from the SNP estimates on the observed Fisher information matrix $I\left(\hat{\theta}\right)$ with respect to $\hat{\theta} \equiv \hat{\theta}(\mathbf{K})$ with $\mathbf{K}$ being the SNP polynomial degrees of the final fit. In particular, for a specific element of $\hat{\theta}$, say $\hat{\theta}_r$, the corresponding (two-sided) Wald-type α- level confidence interval is $\hat{\theta}_r \pm z_{\alpha/2} I\left(\hat{\theta}\right)_{rr}^{-1/2}$, where $I\left(\hat{\theta}\right)_{rr}^{-1/2}$ is the square root of the r^{th} element in the diagonal of the estimated asymptotic variance-covariance matrix $I\left(\hat{\theta}\right)^{-1}$, and $z_{\alpha/2}$ is the $\alpha/2$-quantile of the standard normal distribution. For the pointwise confidence interval of a specific CIF estimate of event type j at a chosen time t, $\widehat{CIF}_j(t)$, I first compute the Wald-type confidence interval for the complementary log-log transform of this quantity, denoted by $c\log\log\left(\widehat{CIF}_j(t)\right)$, using the delta rule:

$$c\log\log\left(\widehat{CIF}_j(t)\right) \pm z_{\alpha/2}\left[\nabla\left(\hat{\theta},t\right)^T I\left(\hat{\theta}\right)^{-1} \nabla\left(\hat{\theta},t\right)\right]^{1/2}$$

where $\nabla\left(\hat{\theta},t\right)$ is the column gradient vector of $c\log\log\left(\widehat{CIF}_j(t)\right)$ as a function of $\hat{\theta}$. Then transforming this confidence interval back to the original CIF scale gives a confidence interval for $\widehat{CIF}_j(t)$. Operating on the cloglog-scale is expected to improve the validity of the normal approximation and avoids having out of range confidence intervals.

Here one might think that using the sandwich type robust variance estimator (see e.g. Equations

(4.5.2) and (4.5.4) in Davison (2008)) for the above calculations would be more appropriate due to a missing rigorous justification for the asymptotic normality for the SNP estimates.. However from all of my simulation results, which will be shown later, conventional and robust standard errors yield similar confidence interval coverage probabilities. In addition theoretical work in linear regression setting shows that sandwich estimators themselves are very variable which can lead to coverage probabilities below nominal values (Kauermann & Carroll (2001)).

3.5 Comparison of cumulative incidence functions based on SNP estimation

The proposed SNP estimates can also be used as the basis for comparing CIFs of a specific event type between two different groups. This approach is a competing risk version of the two-sample test statistic in Doehler & Davidian (2008). The involved null and alternative hypotheses are the same as laid out in Subsection 1.4.2

$$H_0 : CIF_1^1(\cdot) = CIF_1^2(\cdot) \text{ vs.}$$
$$H_A : \left[CIF_1^1(\cdot) \geq CIF_1^2(\cdot) \text{ or } CIF_1^1(\cdot) \leq CIF_1^2\right] \text{ and } CIF_1^1(\cdot) \neq CIF_1^2(\cdot)$$

Here it is assumed without loss of generality that the CIF of the first competing risk is the target for testing. The considered test is also based on an IWD statistic:

$$\int_0^\tau W(t) \left[\widehat{CIF}_1^1(t) - \widehat{CIF}_1^2(t)\right] dt$$

where $W(\cdot)$ is a positive weight function, and τ is usually the minimum of the largest observation times from the compared groups as mentioned in 1.4.2 or a fixed time point within the observed range of total survival time. However, unlike Subsection 1.4.2, here $\widehat{CIF}_1^1$ and $\widehat{CIF}_1^2$ are the SNP estimates instead of the nonparametric ones. The IWD in this setting differs from the IWD in Equation (10) of Doehler & Davidian (2008) in that the survival functions were replaced by the CIFs.

Like other IWD-based tests the weight function $W(\cdot)$ can be manipulated to make the test more sensitive to a specific alternative hypothesis. However, as SNP estimates of the CIFs do not suffer from the issues of instability at time points where the risk set is small like nonparametric estimates, if we are not interested in any specific alternative hypothesis, we can simply set $W \equiv 1$. In fact in a simulation of Doehler & Davidian (2008) that compared two survival functions, the IWD-based tests with $W(\cdot) \equiv 1$ and $W(\cdot) = \hat{C}_1(\cdot)\hat{C}_2(\cdot)\left\{p_1\hat{C}_1(\cdot) + p_2\hat{C}_2(\cdot)\right\}^{-1}$ had similar performance, where $\hat{C}_k(\cdot)$, $k = 1, 2$, are the estimated censoring survival functions for each group.

For gathering statistical evidence to reject or not reject the null hypothesis, as suggested by Doehler

& Davidian (2008) a normal distribution as the null distribution is plausible with the variance of the IWD statistic derived using the delta method. I also make this normality assumption about the null distribution for my test statistic. Specifically, let the asymptotic covariance matrices for the estimated parameters of the fitted competing risks models for groups 1 and 2 be denoted by, respectively, V_1 and V_2. When $W(\cdot)$ is a deterministic function, the variance of the IWD can be estimated using the delta rule as

$$\sigma^2_{IWD} = \nabla IWD\,(\theta_1, \theta_2)^T \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} \nabla IWD\,(\theta_1, \theta_2)$$

where $\nabla IWD\,(\theta_1, \theta_2)$ is the column gradient vector of $IWD\,(\cdot, \cdot)$ as a function of θ_1 and θ_2, the estimates of parameters of the SNP models for groups 1 and 2, respectively. Note that there is no need to obtain information on covariances between parameter estimates from the two SNP models as they are estimated based on independent groups.

Once the IWD statistic and its variance estimate are computed, the p-value can be derived based on comparing the Wald-type statistic $\frac{IWD}{\sigma_{IWD}}$ to the cumulative standard normal distribution. Alternatively, the null distribution and the resulting p-value can be evaluated exactly without relying on a normal approximation by implementing a permutation test or (or Monte-Carlo approximation to it). In Section 3.7, p-values based on the delta method and permutation tests, respectively, are compared for two real data sets.

3.6 Simulation studies

In this Section, I evaluate how well the proposed SNP-based CIF-estimation method performs across a variety of competing risks scenarios with both right and interval censoring. Alternative approaches which were also evaluated are the standard nonparametric estimator of the CIF and CIF estimates from parametric models based on the mixture factorization.

3.6.1 CIF estimation in the presence of right-censoring – simulation set-up

Simulation scenarios

The scenarios in this section consist of competing risks data with right-censoring and I considered both scenarios based on a mixture-factorization and scenarios based on specifying the cause-specific hazard functions. The first 4 scenarios are based on the mixture representation (3.1.1) of the competing risks process. This representation is consistent with the SNP model, flexible and allows for straightforward simulation from a scenario as follows: First, simulate the failure cause D according to a multinomial distribution, then simulate the event time from the corresponding conditional time-to-event time $T \mid$

$D = j$. If the conditional distribution $T|D = j$ was a SNP-distribution, I used a rejection method for simulation which was first suggested by Gallant & Tauchen (1993). A univariate version of this method is detailed in Appendix A.4.

For ease of exposition, I considered only two competing events, $J = 2$, and varied the right-censoring probabilities among 35%, 45% and 65%. Details about how censoring was simulated is discussed in the next subsection. The conditional time-to-event distributions of $T \mid D = j$ and marginal probabilities of failure causes for the simulation study with only right-censoring were varied as summarized in Table 3.1. The first scenario has a Weibull distribution for each conditional time to event distribution i.e. SNP distributions with exponential base densities and $K_1 = K_2 = 0$. The second scenario models one conditional time to event distribution as lognormal and the other with a SNP distribution with $K = 1$ and a standard normal base density. The third scenario has two conditional event times both modelled with SNP distribution with a standard exponential base densities. Finally, the fourth scenario does not involve SNP-distributions but uses two logmixturenormal distributions instead.

As simulations based on mixture factorizations might unfairly favour my own model, the final fifth scenario was based on specifying the cause-specific hazards (CSH) of the competing events instead. A method for simulation from cause-specific hazards is described in Beyersmann et al. (2009). However, in our case it is easier to simulate the data based on two independent latent failure times. In this case, the marginal hazards of the latent times are identical to the cause-specific hazards of the simulated competing risks data. Thus, I first simulated two independent latent failure times according to the distributions specified in the last row of Table 3.1 and then choose $T = \min(T_1, T_2)$ and $D = 1\mathbf{I}(T_1 \leq T_2) + 2\mathbf{I}(T_1 > T_2)$. For this scenario, I considered 45% right-censoring.

For reference, all 5 scenarios with exact parameter choices are detailed in Table 3.1. This table also gives the implied marginal probability of an event of type 1 which for the cause-specific hazards scenario is 0.66.

Table 3.1: Simulation scenarios for CIF estimation under only right-censoring.

Scenario	$f_{T\|D=1}$	$P_1\%$	$f_{T\|D=2}$	$P_2\%$	$\%RC$	t_m
	Mixture representation based scenario					
2 Weibull	$W(1, \exp(-1))$	25	$W(5, \exp(-0.25))$	75	65	50%
2 SNP stdnorm	$\log\mathcal{N}(-1, 0.9^2)$	25	$SNPN(0.1, 0.8, \frac{\pi}{5})$	75	65	50%
2 SNP stdexp	$SNPE(-0.2, 0.8, \frac{\pi}{9})$	50	$SNPE(-0.3, 0.5, -\frac{\pi}{8})$	50	35	90%
2 logmixturenorm	$0.3\log\mathcal{N}(1.2, 0.9^2) + 0.7\log\mathcal{N}(0, 0.6^2)$	50	$0.5\log\mathcal{N}(0, 0.1^2) + 0.5\log\mathcal{N}(1, 0.2^2)$	50	45	75%
	Cause specific hazards based scenario					
Scenario	f_{T_1}	$P_1\%$	f_{T_2}	$P_2\%$	$\%RC$	t_m
Logmixturenorm + Weibull	$0.3\log\mathcal{N}(0.2, 0.36^2) + 0.7\log\mathcal{N}(1.8, 0.36^2)$	66	$W(2.5, \exp(2))$	34	45	75%

Note: $f_{T|D=j}$ is the density of the time to event distribution for event type j conditional on the occurrence of that event type. P_j is the marginal probability of event type j. f_{T_j} is the density of the independent latent failure time distributions associated with event type j used in the cause-specific hazards based scenario. $W(shape, scale)$ means the density of a weibull distribution with a specific shape and scale. $\log\mathcal{N}(\mu, \sigma^2)$ refers to the density of a lognormal distribution with parameters μ and σ. $SNPN(\mu, \sigma, \phi)$ and $SNPE(\mu, \sigma, \phi)$ represent the densities of a random variable T satisfying $\log T = \mu + \sigma Z$; where Z has a SNP distribution with a standard normal base density or e^Z has a SNP distribution with a standard exponential base density, respectively, with spherical coordinates ϕ. %RC is the overall proportion of right-censoring and the values in the column t_m denote the quantile of the marginal distribution of T corresponding to the maximum follow-up duration.

For each scenario, I also considered 2 different sample sizes: $n = 100$ and $n = 500$ leading to a total of 10 simulation settings. For each simulation setting, reported results are based on 200 simulated data sets from that setting.

Simulation of right-censoring

Independent right-censoring was simulated as follows. Let (T, D) be the uncensored data and C the censoring time. Then the observed right-censored data is defined as (T', D') with $T' = \min\{T, C\}$ and $D' = \Delta \times D$, where $\Delta = 1$ if $T \leq C$ and $\Delta = 0$ if $T > C$. The censoring distribution C was simulated as $C = \min\{t_m, C_E\}$, where the maximum follow-up duration t_m was chosen as a suitable quantile of the marginal distribution of T as detailed in Table 3.1 and C_E was simulated according to an exponential distribution with rate λ chosen appropriately to ensure the desired overall right-censoring probability. Specifically, t_m and λ were chosen by simulation based on large competing risks datasets of size $n = 10^6$ from the respective scenarios.

Evaluation criteria

As the simulation study focuses on comparing different methods for CIF-estimation and, moreover, not all scenarios were simulated based on SNP-models, I did not assess the precision of parameter estimates of the SNP-based models. Instead, I used evaluation criteria which directly quantify the precision of the resulting CIF estimates. First I calculated a time independent summary statistic, the so-called average integrated squared error (AISE):

$$AISE_j = \frac{1}{N_S}\sum_{i=1}^{N_S}\left(\int_0^{t_m}\left\{\widehat{CIF}_{ji}(t) - CIF_j(t)\right\}^2 dt\right) \tag{3.6.1}$$

In (3.6.1), N_S is the total number of data sets for a simulation setting. $\widehat{CIF}_{ji}$ refers to the estimated CIF for event type j from the i^{th} data set of the considered simulation setting and CIF_j is the respective true CIF. This measure is an aggregated statistic for the bias of the CIF estimates from time 0 to t_m. Of note, in the CSH (independent latent failure time) scenarios, the true CIF_j is not available in closed form and CIF1 was calculated as follows:

$$\begin{aligned} CIF_1(t) &= P(T \leq t,\ D = 1) \\ &= P(\min(T_1, T_2) \leq t,\ T_1 \leq T_2) \\ &= P(\min(T_1, T_2) \leq t \mid T_1 \leq T_2)P(T_1 \leq T_2) \\ &= P(T_1 \leq t \mid T_1 \leq T_2)P(D = 1) \end{aligned} \tag{3.6.2}$$

where $P(T_1 \leq t \mid T_1 \leq T_2)$ was approximated by the empirical distribution function from an uncensored population level simulated data set of all T_1 with $T_1 \leq T_2$ based on a simulated dataset of size $n = 10^6$. CIF_2 was calculated in the same way.

Second, I calculated point-wise Monte Carlo coverage probabilities of 95% confidence interval for all CIF estimators at two selected time points, t_m and $t_m/2$, based on the complementary log-log transformation. For SNP-based estimators, standard likelihood-based inference was used as detailed in Section 3.4.

Finally I report the median (inter-quartile range) computation time for each method and scenario, where all simulations were conducted on a computer with the following configuration: CPU Intel Core i7-3770 with 8 threads at 3.4GHz, Ram 10G.

Compared estimation methods

SNP models either used a standard exponential or a standard normal based density for both event types; mixed base densities for different event types were not considered. The following information criteria were investigated for choosing both the optimal polynomial degrees and base density: AIC, BICn, BICd, HQCn and HQCd. The resulting best SNP-model was denoted by SB or, to highlight the chosen information criterion, as SB-AIC, SB-BICn, or SB-HQCn.

Information criteria were compared with respect to how frequently the SB model identified both the correct base densities and the correct polynomial degrees K_1 and K_2 . For convenience, I denote the combination of the base density and the polynomial degrees by "base-K". Of note, this comparison was only possible for SNP-based simulation scenarios, i.e. scenarios corresponding to the first 3 rows in Table 3.1. As discussed previously, parametric lognormal or Weibull conditional time-to-event distributions are also considered SNP scenarios as they correspond to SNP-distributions with polynomial degree 0.

To my knowledge, parametric models for competing risks data have not been implemented in standard statistical software. However, my own implementation for SNP-based models can easily fit lognormal (LN) and Weibull (WB) models based on the mixture factorization as a special case and these models were used as parametric benchmarks for my SNP-based methods. The standard nonparametric (NP) estimator of the CIF in the presence of right-censoring has been implemented in many statistical packages. For the simulation study I used the implementation in the R-function `etmCIF` from package `etm` (Allignol et al. (2011)).

3.6.2 CIF estimation in the presence of interval-censoring – simulation set-up

Simulation scenarios

The same simulation scenarios and sample sizes as for the simulation study with right-censoring detailed above were used except that censoring was simulated differently. In particular, the same uncensored data were reused. The choice of the maximum follow-up duration t_m and subsequent right-censoring at t_m was implemented as before. However, rather than simulating additional right-censoring before time t_m, interval-censoring was simulated as follows depending on the chosen quantile of T for the choice of t_m: if t_m corresponds to the 90% quantile of T , I chose 9 time points in $(0, t_m)$ creating 10 equally spaced intervals from 0 to t_m. Likewise when t_m is the median or the 75% quantile of T I chose 4 or 6 time points to get 5 or 7 equally spaced intervals in $[0, t_m]$, respectively. For each subject, their corresponding observation process was defined as occurring at these time point plus some subject-specific "noise" for time points other than 0 and t_m simulated according to a normal

distribution with mean 0 and standard deviation equal to 1/5 of the interval length. If a subject had an event before time t_m then their data was considered interval-censored between the two adjacent time points of the observation process; otherwise they were considered right-censored at t_m.

Nonparametric method for interval-censored competing risks data

For interval-censored data, the used nonparametric estimator of the CIFs is based on the nonparametric maximum likelihood estimator for a bivariate distribution (X, Y), subject to interval-censoring which is described in Gentleman & Vandal (2002) and Maathuis (2003). This estimator had been implemented in function `computeMLE` of the R package `MLEcens` (Maathuis (2013)) which supports interval-censored competing risks data as a special case. Specifically in our setting, X corresponds to the time to event T and Y to the event type D. When T is known to be in an interval $[t_L, t_R]$ and $D = j$, $X = [t_L, t_R]$ and $Y = [j, j]$. When T is right-censored i.e. in $[t, \infty)$, $X = [t, \infty)$ and $Y = [1, J]$, where J is the total number of event types.

As mentioned in Gentleman & Vandal (2002), nonparametric MLE methods for interval-censored data do not always lead to unique estimators and in the competing risks setting, such methods yield lower and upper bounds for the CIF estimates. As the lower and upper bounds were usually very close to each other for my simulation scenarios, I considered the average of both curves as the nonparametic estimate of the CIF for evaluation of the performance of the method. A more specialised way to analyse interval-censored competing risks data has been discussed in Hudgens et al. (2001); however to my knowledge this method is not yet available in R.

3.6.3 CIF estimation in the presence of right-censoring – results

First, I evaluated how frequently SB, i.e. the best SNP-model according to the selected information criterion, chose the correct base-K combination, i.e. the correct base densities and polynomial degrees. Results are reported in Table 3.2. Of note, I found that for the information criteria BIC and HQC, results were very similar whether they were based on the sample size n or the number of events d. Therefore, only results based on n are reported.

Table 3.2: Frequency with which the SB models based on AIC, BICn or HQCn respectively chose the correct base-K for SNP scenarios (first 3 rows). For non-SNP scenarios, the frequency of SB models where the maximal allowed polynomial degree was chosen (i.e. $K_1 = 2$ or $K_2 = 2$) is reported (rows 4 and 5).

Frequency of correct base-K							
		AIC		BICn		HQCn	
Scenario	n	100	500	100	500	100	500
2 Weibull 65% RC		60	123	95	172	75	159
2 SNP stdnorm 65% RC		83	139	118	186	105	173
2 SNP stdexp 35% RC		122	138	153	191	137	174
Frequency of SB with $K_1 = 2$ or $K_2 = 2$							
2 logmixturenorm 45% RC		193	200	135	200	165	200
Logmixturenorm + weibull (CSH) 45% RC		122	199	47	193	90	196

All frequencies are based on 200 simulated data sets per scenario.

Among the SNP scenarios with smaller sample size ($n = 100$), the proportion that SB chose the correct base-K is quite low, ranging from 44.2% for AIC to 61% for BICn. For larger sample size ($n = 500$), these proportions increased considerably and ranged from 66.7% for AIC to 91.5% for BICn with an intermediate value of 84.3% for HQCn. As a comparison, Doehler & Davidian (2008) reported proportions ranging from 84% to 90% for HQCn to correctly identify $K = 0$ for parametric lognormal or Weibull models in survival analysis based on a sample size of 400. This is similar but somewhat higher than my results which is not unexpected because if there are multiple event types, it is more difficult to identify the correct base-K for all of them.

Several papers (Doehler & Davidian (2008) and Zhang & Davidian (2008)) suggested that one should choose HQC as the most appropriate information criterion. However, the results from Table 3.2 demonstrate that BICn is considerably more successful in identifying the correct base-Ks than the other criteria. Thus on this basis, my results suggest that BIC should be used. However, the last 2 rows in Table 3.2 shows the frequency that SB chose complex SNP with $K_1 = 2$ or $K_2 = 2$ for non-SNP scenarios. As can be seen, AIC tends to fit more complex models in this situation and this might indicate that AIC is more successful in approximating non-SNP distributions. Indeed, as shown below, in terms of AISE and pointwise 95%-CI coverage in non-SNP scenarios, AIC generally leads to the best results.

Table 3.3 shows relative average integrated square error (AISE) of parametric and nonparametric methods compared to the SB-AIC model. If a parametric Weibull or log-normal model was the true model for the respective CIF, SB-AIC performed almost as well as the true parametric model. Of note, Figure 3.6.1 shows that within $[0, t_m]$, the true CIF_1 of the 2 logmixturenorm scenario and the true CIF_2 of the CSH scenario should be estimated easily by one of the compared parametric methods. If

this was not the case, SB-AIC usually outperformed the parametric models, sometimes dramatically so. SB-AIC consistently outperformed the nonparametric model for the first 3 simulation scenarios which were based on true underlying SNP models. However, for the non-SNP scenarios, the situation was less clear with advantages of either method for different scenarios. Nonetheless, from Figure 3.6.1, my SNP method on average produced quite good CIF estimates for these scenarios.

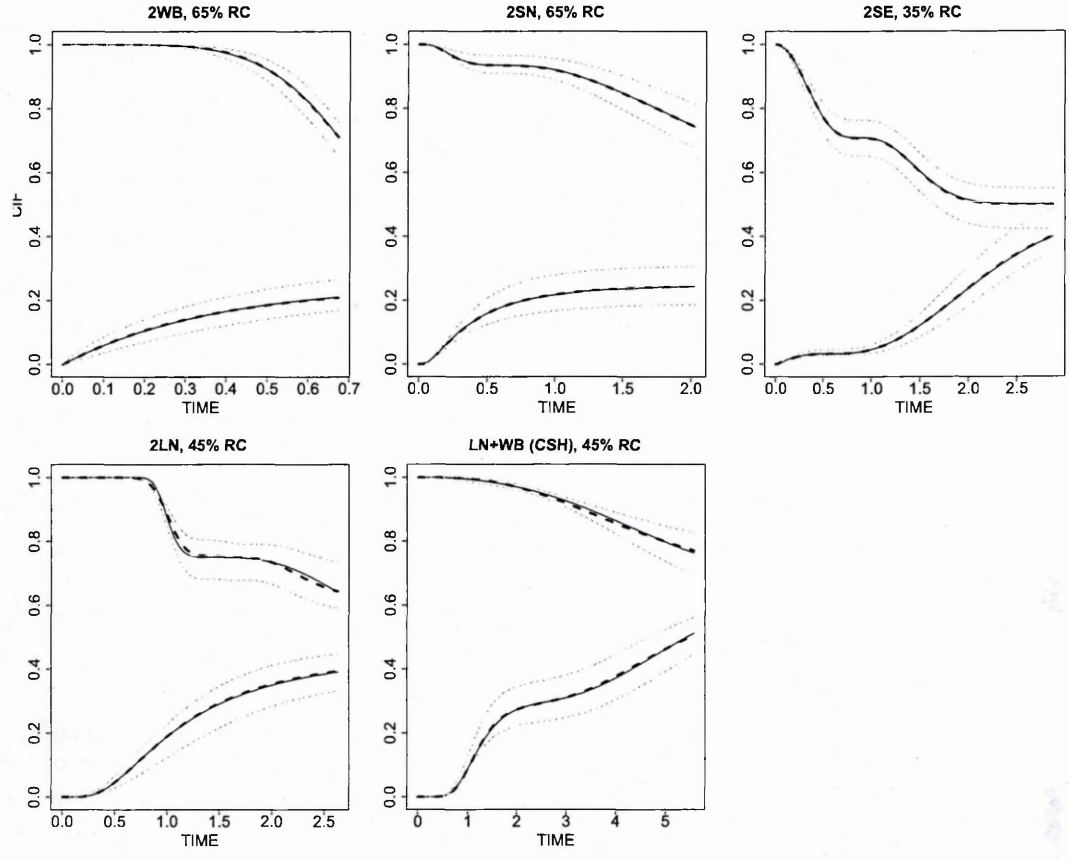
Compared to SB-AIC, SB-BICn usually lead to smaller AISE for SNP-based scenarios but worse results for non-SNP scenarios (Table 3.3). Conclusions for SB-HQCn compared to SB-AIC were the same though in this case AISE estimates were closer to each other (data not shown). Figure 3.6.1 visualizes estimated CIFs for the SB-AIC model with $n = 500$ and nicely illustrates that the SNP-approach is capable to closely approximate quite complex underlying CIF functions.

Table 3.3: Average integrated square error (AISE) for different estimation methods for all scenarios with right-censoring. Shown is the relative performance (with standard error) of parametric and nonparametric methods versus the SB-AIC model, respectively, and AISE values for the SB-AIC and SB-BICn model.

	$n = 100$		$n = 500$	
Scenario	CIF_1	CIF_2	CIF_1	CIF_2
2 Weibull 65% RC				
LN/SB-AIC	**0.94** 0.014	**1.08** 0.030	**1.00** 0.026	**1.55** 0.080
WB/SB-AIC	**1.01** 0.011	**0.91** 0.018	**1.00** 0.010	**0.96** 0.029
NP/SB-AIC	**1.09** 0.015	**1.13** 0.023	**1.11** 0.014	**1.20** 0.033
SB-AIC $\times 10^4$	8.90	4.50	1.62	0.74
SB-BICn $\times 10^4$	8.73	4.37	1.62	0.72
2 SNP stdnorm 65% RC				
LN/SB-AIC	**0.96** 0.009	**1.42** 0.060	**0.99** 0.004	**3.90** 0.289
WB/SB-AIC	**0.99** 0.007	**1.22** 0.043	**1.08** 0.011	**2.79** 0.189
NP/SB-AIC	**1.02** 0.010	**1.13** 0.024	**1.05** 0.008	**1.18** 0.029
SB-AIC $\times 10^4$	37.35	23.85	8.05	4.88
SB-BICn $\times 10^4$	36.96	23.26	8.01	4.79
2 SNP stdexp 35% RC				
LN/SB-AIC	**2.81** 0.176	**1.24** 0.027	**12.61** 1.183	**2.85** 0.152
WB/SB-AIC	**1.54** 0.063	**1.21** 0.021	**3.70** 0.288	**2.63** 0.132
NP/SB-AIC	**1.17** 0.024	**1.04** 0.006	**1.27** 0.034	**1.06** 0.006
SB-AIC $\times 10^4$	41.61	76.05	7.00	12.96
SB-BICn $\times 10^4$	40.92	76.02	6.89	12.88
2 logmixturenorm 45% RC				
LN/SB	**1.01** 0.018	**1.73** 0.067	**1.48** 0.059	**4.76** 0.305
WB/SB	**1.09** 0.024	**1.95** 0.095	**1.44** 0.072	**6.89** 0.477
NP/SB	**1.05** 0.012	**0.89** 0.015	**1.07** 0.011	**0.85** 0.016
SB-AIC $\times 10^4$	49.15	49.47	10.38	10.45
SB-BICn $\times 10^4$	50.31	55.14	11.39	10.47
Logmixturenorm + weibull (CSH) 45% RC				
LN/SB-AIC	**1.40** 0.054	**0.98** 0.018	**4.15** 0.333	**1.12** 0.032
WB/SB-AIC	**1.45** 0.041	**0.91** 0.017	**4.26** 0.333	**0.76** 0.030
NP/SB-AIC	**1.04** 0.014	**1.12** 0.023	**1.06** 0.013	**0.94** 0.028
SB-AIC $\times 10^4$	122.09	55.96	22.91	10.28
SB-BICn $\times 10^4$	133.74	55.32	23.85	10.69

Note: LN/SB-AIC, WB/SB-AIC and NP/SB-AIC are respectively the ratios (in bold) of the AISE of the parametric lognormal, Weibull and the nonparametric models, respectively, versus the SB-AIC model with corresponding bootstrap standard errors. For each scenario, the last 2 rows give AISE values for SB-AIC and SB-BICn.

Figure 3.6.1: Simulation results for SB-AIC models for all 5 scenarios with right-censoring and a sample size of $n = 500$.



Bold lines show the true CIF_1 for event type 1, and $1 - CIF_2$ for event type 2. Bold dashed lines show the corresponding point-wise averaged fitted CIFs across 200 simulation runs. Light dashed lines show curves resulting from the two simulation runs leading to the minimal and maximum average residual from the true curve based on 300 equally spaced time point from 0 to t_m, i.e. they display the worst observed under- and over-estimation. From left to right, top to bottom are the scenarios: 2 weibull, 2 SNP stdnorm, 2 SNP stdexp, 2 logmixturenormal and logmixturenormal + Weibull.

Pointwise Monte Carlo coverage probabilities of the nominal 95% confidence intervals (95% CI) are displayed in Table 3.4. They were based on 200 simulation runs for each scenario which implies that the Monte Carlo standard error of the coverage probability estimate is 1.5%. Parametric estimators performed well if the true model was the same parametric model but showed dramatic under-coverage in several other situations. The SB-AIC model achieved observed coverage $\geq$90% for all scenarios except for the estimation of CIF_1 at $0.5t_m$ in the last scenario with $n = 500$ where observed coverage was only 82.5%. Also, there was some evidence of less dramatic undercoverage (observed coverage <92%) in a few other scenarios. SB-BIC performed similarly to SB-AIC in SNP scenarios but worse in non-SNP scenarios. The nonparametric estimator performed well with observed coverage $\geq$92% across all scenarios except for one very low observed coverage for the first

scenario with $n = 100$ which may have occurred because none of the simulated data sets from this scenario has more than one event of type 2 before time $0.5t_m$.

Table 3.4: Observed coverage probabilities of nominal 95% CI for the CIFs at $t_m/2$ and t_m in all scenarios.

	$n = 100$				$n = 500$			
Time point	$0.5t_m$		t_m		$0.5t_m$		t_m	
Scenario	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2
2 Weibull 65% RC								
LN	95.0	93.0	96.5	95.5	96.0	79.5	95.0	94.0
WB	94.5	95.5	96.0	94.5	95.0	96.5	98.0	94.5
NP	93.0	59.5	96.5	94.5	95.5	95.0	98.0	95.0
SB-AIC	92.5	94.0	96.5	94.5	95.5	91.0	98.0	94.0
SB-BICn	94.5	95.0	96.5	95.0	95.0	93.0	98.0	94.0
2 SNP stdnorm 65% RC								
LN	94.0	65.5	95.5	94.0	92.0	8.5	93.0	77.0
WB	94.5	76.5	96.0	95.5	90.5	23.5	93.0	90.0
NP	95.5	96.5	95.0	95.0	92.0	95.5	92.5	92.0
SB-AIC	94.0	92.5	95.5	94.0	92.0	92.0	93.0	91.0
SB-BICn	94.0	92.0	95.5	93.5	91.5	91.5	93.0	92.0
2 SNP stdexp 35% RC								
LN	40.5	86.5	63.0	95.0	0.0	65.5	1.0	94.5
WB	65.0	86.5	92.0	95.0	19.5	61.0	82.0	96.5
NP	95.5	94.5	96.5	94.5	97.0	95.0	96.0	96.5
SB-AIC	92.5	92.0	95.0	94.5	96.5	94.0	96.0	96.5
SB-BICn	92.0	91.5	94.5	94.5	96.5	94.5	96.5	96.5
2 logmixturenorm 45% RC								
LN	90.5	66.5	91.0	92.0	79.0	10.0	77.5	89.5
WB	88.5	33.5	93.5	92.0	81.5	1.5	91.5	89.5
NP	95.0	93.5	95.5	95.5	93.5	94.5	93.5	92.0
SB-AIC	92.0	90.5	94.5	93.0	93.0	93.0	93.0	92.5
SB-BICn	90.0	83.0	93.5	93.0	89.5	93.5	91.5	92.5
Logmixturenorm + weibull (CSH) 45% RC								
LN	89.5	87.0	95.0	94.0	75.0	74.0	93.5	95.0
WB	93.5	91.5	95.5	93.0	93.0	97.0	95.0	95.5
NP	95.5	94.5	94.5	93.0	93.5	95.0	94.5	95.5
SB-AIC	94.5	90.0	94.0	93.0	93.5	82.5	94.5	94.5
SB-BICn	93.0	88.5	95.0	93.0	92.0	80.5	95.0	95.5

Note: LN and WB are the Monte Carlo coverage probabilities of nominal 95% Wald confidence intervals (CIs) based on the parametric 2 x lognormal and 2 x weibull models respectively. SB-AIC and SB-BICn are the observed coverage probabilities of the best SNP models based on AIC and BICn respectively. NP is the observed coverage probability of the nonparametric model. The 95% CIs were calculated based on the cloglog transforms of the CIFs. Estimated standard error of Monte Carlo coverage entries $\approx 1.5\%$.

In conclusion, regarding the choice of the most appropriate information criterion, the simulation results suggest that BICn or HQCn are preferable if one thinks that it is likely that the true data-generating mechanism is close to a Weibull or lognormal distribution or a SNP-distribution with a low

degree polynomial. However, for non-SNP scenarios, AIC generally leads to lower AISE and better coverage of associated CI. Regarding the observed undercoverage of SNP-based methods in some instances, bootstrap methods might provide improved results. Indeed, Doehler & Davidian (2008) showed that this is the case in the survival setting. However, in our setting computing times as tabulated in Table 3.5 are considerable and this prevented me from further exploring the bootstrap in a simulation setting.

Table 3.5: Median and IQR of the computing time (in second) for determining the best SNP model for a standard normal (SNP-stdnorm) or exponential base density (SNP-stdexp), respectively, based on AIC.

	SNP-stdnorm		SNP-stdexp	
Scenario	$n = 100$	$n = 500$	$n = 100$	$n = 500$
2 Weibull 65% RC				
	16.5 (9.9, 34.1)	52.0 (26.8, 96.0)	16.6 (11.9, 39.8)	44.4 (36.5, 102.3)
2 SNP stdnorm 65% RC				
	21.2 (15.0, 37.4)	54.3 (48.7, 93.7)	42.2 (15.1, 57.7)	127.9 (71.8, 151.1)
2 SNP stdexp 35% RC				
	25.6 (16.9, 31.9)	71.0 (55.1, 82.0)	39.4 (31.4, 51.3)	97.1 (87.4, 122.1)
2 logmixturenorm 45% RC				
	22.2 (9.0, 34.5)	76.1 (69.7, 94.7)	30.0 (22.6, 37.1)	78.7 (72.9, 94.6)
Logmixturenorm + weibull (CSH) 45% RC				
	18.3 (9.1, 33.1)	67.2 (59.4, 88.5)	24.0 (9.3, 33.8)	68.9 (63.8, 91.6)

In addition I observed that in non-SNP scenarios, even the conservative SB-BICn frequently chose $K_1 = 2$ or $K_2 = 2$. This suggests that more complex SNP models may be needed. Therefore, I repeated the entire simulation study with $K_{max} = 3$. Table 3.6 indicates that indeed $K_1 = 3$ or $K_2 = 3$ is frequently chosen if I allow it but computing time also increases substantially (Table 3.7). A comparison of AISE (Table 3.8) and coverage probabilities of nominal 95% CI (Table 3.9) between $K_{max} = 2$ and $K_{max} = 3$ indicates that there is indeed some improvement in the non-SNP scenarios but that the gains are only moderate. In view of the increased computing time and only limited improvements, I consider only $K_{max} = 2$ for all subsequent simulations. However, these results may indicate that if more efficient algorithms or higher computing power were available, further gains might be possible by not restricting K_{max} at all and only letting the information criterion decide on the appropriate complexity.

Table 3.6: Number of best SNP models based on AIC, BICn and HQCn allowing for $K_{max} = 3$ which chose $K_1 = 3$ and/or $K_2 = 3$. Total number of simulation runs was 200 data sets per scenario.

		AIC		BICn		HQCn	
Scenario	n	100	500	100	500	100	500
2 Weibull 65% RC		5	11	0	0	1	1
2 SNP stdnorm 65% RC		18	15	0	2	2	3
2 SNP stdexp 35% RC		22	22	0	0	8	8
2 logmixturenorm 45% RC		167	198	74	179	118	196
Logmixturenorm + weibull (CSH) 45% RC		71	108	14	66	44	73

Table 3.7: Median and IQR of the computing time (in second) for choosing the best SNP model for a standard normal (SNP-stdnorm) or exponential base density (SNP-stdexp), respectively, based on AIC with $K_{max} = 3$.

	SNP-stdnorm		SNP-stdexp	
Scenario	$n = 100$	$n = 500$	$n = 100$	$n = 500$
2 Weibull 65% RC	16.5 (9.9, 63.3)	52.0 (26.8, 234.7)	16.6 (11.9, 50.9)	44.4 (36.5, 202.2)
2 SNP stdnorm 65% RC	21.2 (15.0, 105.4)	54.3 (48.7, 231.6)	44.5 (15.1, 146.7)	297.5 (71.8, 598.3)
2 SNP stdexp 35% RC	60.6 (16.9, 149.6)	183.6 (155.8, 430.5)	39.8 (31.9, 129.2)	97.1 (87.4, 302.9)
2 logmixturenorm 45% RC	58.7 (9.0, 96.5)	231.3 (179.5, 427.4)	73.8 (58.1, 123.7)	206.8 (187.9, 293.8)
Logmixturenorm + weibull (CSH) 45% RC	18.3 (9.1, 107.0)	209.0 (164.3, 551.0)	68.4 (9.3, 119.1)	211.4 (190.1, 248.8)

Table 3.8: Comparison of AISE between best SNP models with $K_{max} = 2$ and $K_{max} = 3$.

	$n = 100$		$n = 500$	
Scenario	CIF_1	CIF_2	CIF_1	CIF_2
2 Weibull 65% RC				
SB-AIC $K_{max} = 2/K_{max} = 3$	**1.00** 0.003	**1.00** 0.000	**1.00** 0.006	**1.00** 0.006
SB-BICn $K_{max} = 2/K_{max} = 3$	**1.00** 0.000	**1.00** 0.000	**1.00** 0.000	**1.00** 0.000
2 SNP stdnorm 65% RC				
SB-AIC $K_{max} = 2/K_{max} = 3$	**1.00** 0.001	**0.99** 0.005	**1.00** 0.001	**0.98** 0.014
SB-BICn $K_{max} = 2/K_{max} = 3$	**1.00** 0.000	**1.00** 0.000	**1.00** 0.000	**1.00** 0.001
2 SNP stdexp 35% RC				
SB-AIC $K_{max} = 2/K_{max} = 3$	**1.00** 0.006	**0.99** 0.003	**0.99** 0.008	**1.00** 0.002
SB-BICn $K_{max} = 2/K_{max} = 3$	**1.00** 0.000	**1.00** 0.000	**1.00** 0.000	**1.00** 0.000
2 logmixturenorm 45% RC				
SB-AIC $K_{max} = 2/K_{max} = 3$	**1.01** 0.007	**1.07** 0.010	**1.00** 0.007	**1.13** 0.015
NP / SB-AIC $K_{max} = 3$	**1.06** 0.012	**0.95** 0.010	**1.07** 0.011	**0.97** 0.008
SB-BICn $K_{max} = 2/K_{max} = 3$	**1.00** 0.006	**1.04** 0.008	**0.99** 0.010	**1.10** 0.015
NP / SB-BICn $K_{max} = 3$	**1.03** 0.017	**0.83** 0.023	**0.97** 0.017	**0.94** 0.012
Logmixturenorm + weibull (CSH) 45% RC				
SB-AIC $K_{max} = 2/K_{max} = 3$	**1.01** 0.008	**1.01** 0.008	**0.99** 0.009	**1.09** 0.020
NP / SB-AIC $K_{max} = 3$	**1.05** 0.013	**1.12** 0.024	**1.05** 0.012	**1.02** 0.029
SB-BICn $K_{max} = 2/K_{max} = 3$	**1.02** 0.014	**1.00** 0.003	**1.02** 0.009	**1.08** 0.024
NP / SB-BICn $K_{max} = 3$	**0.96** 0.022	**1.13** 0.024	**1.04** 0.011	**0.98** 0.034

Note: SB-AIC $K_{max} = 2/K_{max} = 3$ is the ratio (in bold) of the AISE from the best AIC-based SNP models using $K_{max} = 2$ to using $K_{max} = 3$ with the corresponding bootstrap standard errors. The same is for SB-BICn. In non-SNP scenarios, NP / SB-AIC $K_{max} = 3$ and NP / SB-BICn $K_{max} = 3$ are, respectively, the relative AISE of the nonparametric estimates versus the best SNP estimates using AIC and BICn with $K_{max} = 3$.

Table 3.9: Observed coverage probabilities of nominal 95% CI for CIFs estimation using $K_{max} = 2$ and $K_{max} = 3$ at $t_m/2$ and t_m.

	$n = 100$				$n = 500$			
Time point	$0.5t_m$		t_m		$0.5t_m$		t_m	
Scenario	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2
2 Weibull 65% RC								
SB-AIC $K_{max} = 2$	92.5	94.0	96.5	94.5	95.5	91.0	98.0	94.0
SB-AIC $K_{max} = 3$	92.5	94.0	96.5	94.5	95.5	91.0	98.0	94.0
SB-BICn $K_{max} = 2$	94.5	95.0	96.5	95.0	95.0	93.0	98.0	94.0
SB-BICn $K_{max} = 3$	94.5	95.0	96.5	95.0	95.0	93.0	98.0	94.0
2 SNP stdnorm 65% RC								
SB-AIC $K_{max} = 2$	94.0	92.5	95.5	94.0	92.0	92.0	93.0	91.0
SB-AIC $K_{max} = 3$	94.0	92.5	95.5	93.5	91.5	92.5	93.0	91.0
SB-BICn $K_{max} = 2$	94.0	92.0	95.5	93.5	91.5	91.5	93.0	92.0
SB-BICn $K_{max} = 3$	94.0	92.0	95.5	93.5	91.5	91.5	93.0	92.0
2 SNP stdexp 35% RC								
SB-AIC $K_{max} = 2$	92.5	92.0	95.0	94.5	96.5	94.0	96.0	96.5
SB-AIC $K_{max} = 3$	92.0	92.0	95.5	94.5	96.5	94.0	96.0	96.5
SB-BICn $K_{max} = 2$	92.0	91.5	94.5	94.5	96.5	94.5	96.5	96.5
SB-BICn $K_{max} = 3$	92.0	91.5	94.5	94.5	96.5	94.5	96.5	96.5
2 logmixturenorm 45% RC								
SB-AIC $K_{max} = 2$	92.0	90.5	94.5	93.0	93.0	93.0	93.0	92.5
SB-AIC $K_{max} = 3$	91.5	91.5	95.0	93.5	93.0	94.0	92.5	92.0
SB-BICn $K_{max} = 2$	90.0	83.0	93.5	93.0	89.5	93.5	91.5	92.5
SB-BICn $K_{max} = 3$	90.0	82.5	93.0	93.5	90.0	93.0	91.5	91.5
Logmixturenorm + weibull (CSH) 45% RC								
SB-AIC $K_{max} = 2$	94.5	90.0	94.0	93.0	93.5	82.5	94.5	94.5
SB-AIC $K_{max} = 3$	95.0	91.0	94.0	93.0	94.0	86.5	96.0	95.5
SB-BICn $K_{max} = 2$	93.0	88.5	95.0	93.0	92.0	80.5	95.0	95.5
SB-BICn $K_{max} = 3$	93.5	88.5	94.5	93.0	93.5	83.5	96.0	95.5

Note: SB-AIC $K_{max} = 2$ and SB-BICn $K_{max} = 2$ are, respectively, the Monte Carlo coverage probabilities of nominal 95% CI for SB-AIC and SB-BICn using $K_{max} = 2$. The same is for $K_{max} = 3$. The 95% CIs were calculated based on the cloglog transforms of the CIFs. Estimated standard error of Monte Carlo coverage entries $\approx 1.5\%$.

3.6.4 CIF estimation of two competing risks in the presence of interval and right-censoring – results

Similarly to the right-censored case, Table 3.10 demonstrates that the highest allowed SNP polynomial degree of 2 is frequently reached for non-SNP scenarios in the presence of interval censoring. However, I observed much lower frequency of correct base-K; especially for low sample size. Despite this, SB-AIC in general produced good CIF estimates (Figure 3.6.2) and outperformed the nonparametric estimator across all scenarios in terms of AISE (Table 3.12). SB-AIC also outperformed parametric estimators in most settings if the parametric model did not reflect the true data-generating mechan-

ism. Observed pointwise coverage of SB-AIC models was mostly close to 95% but clearly below 90% in 3 instances (2 Weibull scenario: CIF_2 at $0.5t_m$, $n = 100$ and $n = 500$; CSH scenario: CIF_2 at $0.5t_m$, $n = 500$). Of note, I did not calculate confidence intervals for nonparametric estimates as they are not available in standard software and may have non-standard asymptotic properties (Maathuis (2006)). With respect to computing time, somewhat surprisingly, my implementation ran faster for interval-censored than for right-censored data, see Table 3.11. Finally, as in the simulation with only right-censoring, I expect that simulation results for SNP-based models in this Section may be improved for higher K_{max}.

Table 3.10: Frequency with which the SB models based on AIC, BICn or HQCn respectively chose the correct base-K for SNP scenarios (first 3 rows). For non-SNP scenarios, the frequency with of SB models with $K_1 = 2$ or $K_2 = 2$ is reported (rows 4 and 5).

Frequency of correct base-K							
		AIC		BICn		HQCn	
Scenario	n	100	500	100	500	100	500
2 Weibull 65% RC		33	58	94	143	60	103
2 SNP stdnorm 65% RC		40	93	72	168	65	132
2 SNP stdexp 35% RC		86	114	130	193	112	162
Frequency of SB with $K_1 = 2$ or $K_2 = 2$							
2 logmixturenorm 45% RC		191	200	139	200	171	200
Logmixturenorm + weibull (CSH) 45% RC		98	179	15	104	55	153

All frequencies are based on 200 simulated data sets per scenario.

Table 3.11: Average and IQR of the performance time (in second) of AIC-based SNP-stdnorm and SNP-stdexp methods.

	SNP-stdnorm		SNP-stdexp	
Scenario	$n = 100$	$n = 500$	$n = 100$	$n = 500$
2 Weibull 65% RC				
	17.8 (8.9, 31.2)	42.0 (20.6, 67.8)	17.5 (9.0, 30.8)	40.0 (22.3, 73.0)
2 SNP stdnorm 65% RC				
	21.8 (14.6, 31.9)	53.6 (37.7, 83.4)	21.9 (9.6, 33.4)	62.6 (38.9, 86.7)
2 SNP stdexp 35% RC				
	27.1 (17.2, 32.7)	71.9 (60.4, 81.2)	27.4 (18.1, 34.8)	67.0 (56.8, 80.6)
2 logmixturenorm 45% RC				
	22.4 (17.5, 31.5)	60.7 (45.7, 68.9)	25.9 (18.6, 33.6)	60.4 (55.9, 65.9)
Logmixturenorm + weibull (CSH) 45% RC				
	22.9 (14.6, 32.5)	58.7 (42.4, 76.8)	24.1 (15.3, 37.6)	54.4 (45.6, 69.0)

Table 3.12: Average integrated square error (AISE) for different estimation methods for all scenarios with interval-censoring. Shown is the relative performance (with standard error) of parametric and nonparametric methods versus the SB-AIC model, respectively, and AISE values for the SB-AIC and SB-BICn model.

	$n = 100$		$n = 500$	
Scenario	CIF_1	CIF_2	CIF_1	CIF_2
2 Weibull 65% RC				
LN/SB-AIC	**0.94** 0.013	**0.91** 0.042	**0.92** 0.019	**0.58** 0.052
WB/SB-AIC	**0.96** 0.013	**0.85** 0.031	**0.91** 0.016	**0.37** 0.032
NP/SB-AIC	**1.31** 0.027	**1.86** 0.096	**1.70** 0.061	**3.24** 0.193
SB-AIC $\times 10^4$	8.11	5.88	1.76	1.73
SB-BICn $\times 10^4$	7.82	5.09	1.66	0.82
2 SNP stdnorm 65% RC				
LN/SB-AIC	**0.95** 0.009	**1.47** 0.063	**0.85** 0.018	**3.53** 0.223
WB/SB-AIC	**0.95** 0.010	**1.32** 0.052	**0.87** 0.021	**2.78** 0.167
NP/SB-AIC	**1.26** 0.025	**1.53** 0.053	**1.58** 0.050	**2.06** 0.109
SB-AIC $\times 10^4$	36.96	23.14	8.67	5.54
SB-BICn $\times 10^4$	35.80	23.77	7.80	5.29
2 SNP stdexp 35% RC				
LN/SB-AIC	**2.03** 0.121	**1.18** 0.025	**6.50** 0.517	**2.67** 0.139
WB/SB-AIC	**1.29** 0.047	**1.13** 0.018	**2.48** 0.153	**2.29** 0.104
NP/SB-AIC	**1.37** 0.039	**1.16** 0.017	**1.68** 0.067	**1.38** 0.033
SB-AIC $\times 10^4$	38.12	78.49	8.26	13.53
SB-BICn $\times 10^4$	37.34	77.49	7.95	13.12
2 logmixturenorm 45% RC				
LN/SB-AIC	**0.96** 0.021	**1.79** 0.079	**1.44** 0.058	**5.02** 0.327
WB/SB-AIC	**1.08** 0.024	**2.04** 0.104	**1.65** 0.085	**7.12** 0.490
NP/SB-AIC	**1.25** 0.025	**1.25** 0.027	**1.52** 0.047	**1.46** 0.048
SB-AIC $\times 10^4$	50.28	47.14	11.14	9.45
SB-BICn $\times 10^4$	49.12	48.13	12.24	9.68
Logmixturenorm + weibull (CSH) 45% RC				
LN/SB-AIC	**1.32** 0.054	**0.93** 0.023	**3.48** 0.210	**1.21** 0.035
WB/SB-AIC	**1.37** 0.045	**0.88** 0.018	**3.55** 0.206	**0.85** 0.026
NP/SB-AIC	**1.33** 0.028	**1.33** 0.037	**1.57** 0.058	**1.63** 0.074
SB-AIC $\times 10^4$	117.32	49.85	25.88	10.21
SB-BICn $\times 10^4$	120.14	46.82	26.59	10.42

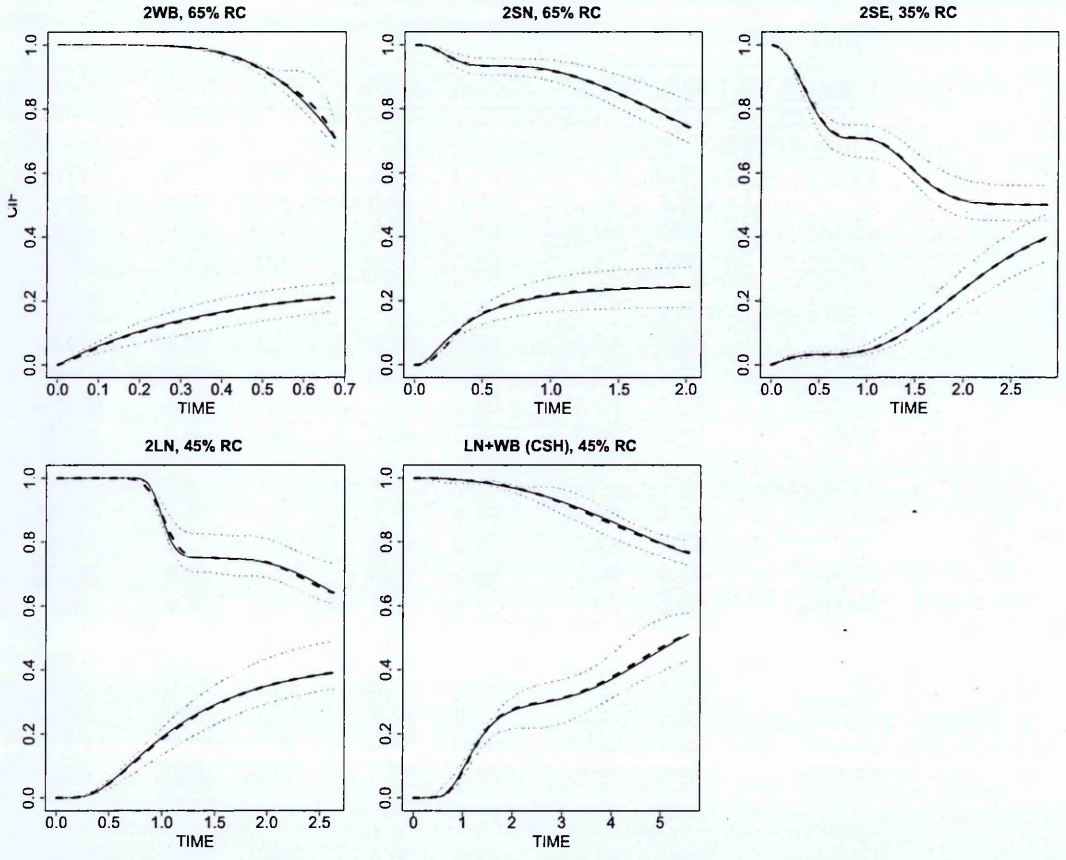
Note: LN/SB-AIC, WB/SB-AIC and NP/SB-AIC are respectively the ratios (in bold) of the AISE of the parametric lognormal, Weibull and the nonparametric models, respectively, versus the SB-AIC model with corresponding bootstrap standard errors. For each scenario, the last 2 rows give AISE values for SB-AIC and SB-BICn.

Table 3.13: Observed coverage probabilities of nominal 95% CI for the CIFs at $t_m/2$ and t_m in all scenarios.

	$n = 100$				$n = 500$			
Time point	$0.5t_m$		t_m		$0.5t_m$		t_m	
Scenario	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2
2 Weibull 65% RC								
LN	96.0	84.0	95.0	92.5	94.5	91.0	97.5	95.5
WB	95.5	86.5	95.5	93.0	95.0	97.5	97.0	97.5
SB-AIC	95.0	80.0	95.5	92.5	94.5	84.0	97.5	97.5
SB-BICn	96.0	85.0	95.5	93.0	94.5	91.0	97.0	96.5
2 SNP stdnorm 65% RC								
LN	94.5	69.0	95.0	94.0	93.0	10.0	94.0	79.5
WB	95.0	80.5	96.0	94.5	94.5	22.5	94.0	90.0
SB-AIC	96.0	95.5	95.0	94.0	92.0	93.5	94.0	91.0
SB-BICn	94.5	93.5	95.5	94.0	92.5	92.5	94.0	91.5
2 SNP stdexp 35% RC								
LN	50.5	88.5	84.5	95.5	3.0	72.5	15.5	94.5
WB	77.5	89.5	96.0	94.5	45.5	76.0	86.5	95.5
SB-AIC	94.0	94.5	94.5	94.5	93.0	93.5	95.5	95.5
SB-BICn	95.5	94.5	95.0	94.5	94.0	94.0	95.0	95.5
2 logmixturenorm 45% RC								
LN	93.5	72.0	93.5	90.0	69.5	23.0	88.5	97.5
WB	90.0	45.0	93.0	92.5	70.5	0.5	94.0	92.5
SB-AIC	93.0	94.5	94.0	93.0	92.5	97.0	96.0	98.0
SB-BICn	93.5	93.0	93.0	92.5	87.5	97.0	95.0	97.0
Logmixturenorm + weibull (CSH) 45% RC								
LN	92.5	92.0	95.5	96.0	76.0	78.0	92.5	96.0
WB	97.0	94.5	95.5	95.0	93.5	93.0	94.5	96.5
SB-AIC	97.0	93.0	96.0	96.0	94.0	86.5	96.0	97.0
SB-BICn	97.0	92.0	96.0	95.0	92.0	84.0	95.5	97.0

Note: LN and WB are the Monte Carlo coverage probabilities of nominal 95% Wald confidence intervals (CIs) based on the parametric 2 x lognormal and 2 x weibull models respectively. SB-AIC and SB-BICn are the observed coverage probabilities of the best SNP models based on AIC and BICn respectively. NP is the observed coverage probability of the nonparametric model. The 95% CIs were calculated based on the cloglog transforms of the CIFs. Estimated standard error of Monte Carlo coverage entries $\approx 1.5\%$.

Figure 3.6.2: Simulation results for SB-AIC models for all 5 scenarios with interval-censoring and a sample size of $n = 500$.



Bold lines show the true CIF_1 for event type 1, and $1 - CIF_2$ for event type 2. Bold dashed lines show the corresponding point-wise averaged fitted CIFs across 200 simulation runs. Light dashed lines show curves resulting from the two simulation runs leading to the minimal and maximum average residual from the true curve based on 300 equally spaced time point from 0 to t_m, i.e. they display the worst observed under- and over-estimation. From left to right, top to bottom are the scenarios: 2 weibull, 2 SNP stdnorm, 2 SNP stdexp, 2 logmixturenormal and logmixturenormal + Weibull.

3.7 Application

In this section, the proposed SNP estimator for the CIF is applied to several data sets from clinical studies conducted at the Oxford University Clinical Research Unit in Viet Nam (OUCRU-VN) and publicly available data sets. In all applications, I report the SB-AIC estimator used in the previous simulations, i.e. the best SNP model selected according to AIC, whose K_{max} is set at 3. For data with only right-censoring, I visually compare my SB-AIC estimates of the CIFs to those given by the nonparametric method described in Subsection 3.6.1. For interval-censored data, I use the nonparametric approach mentioned in Subsection 3.6.2.

3.7.1 Initiation of antiretroviral therapy (ART) in HIV–associated tuberculous meningitis (TBM)

This data set is from a recent randomized control trial (RCT) conducted by OUCRU-VN which compared immediate versus delayed antiretroviral therapy (ART) initiation in HIV-positive patients with TBM (Török et al. (2011)). The primary endpoint of the trial was overall survival. The clinical study did not detect a significant mortality difference between the two groups but observed significantly more severe (grade 4) adverse events in the immediate ART group supporting delayed initiation of ART .

Here, I summarize two secondary competing risks endpoints of the study: First, the time to the first neurological event (harmful event) or prior death i.e. death before experiencing any neurological event (harmful event) and second, in the subset of patients with a Glasgow coma score <15 at enrolment, the time to coma clearance (beneficial event) or prior death (harmful event). For the precise definitions of these outcomes, I refer to Section "Outcome Assessment" in Török et al. (2011). Patients without an event were right-censored at their last follow-up visit which was scheduled to occur at 12 months of follow up and this led to 33% right-censored observations for the first endpoint (neurological event or prior death) and 5% for the second (coma clearance or death).

For the outcome of the time to first neurological event or death, 101/253 patients experienced a neurological event and 68 died without a prior neurological event. Figure 3.7.1 displays the estimated cumulative incidence functions of the time to the first neurological event or prior death by treatment arm in all randomized patients and stratified by severity of the patient as quantified by the TBM grade at enrolment. In all displays, SNP CIF estimates closely agree with the nonparametric estimates. For 6 out of 8 fitted SB-AIC models, standard normal base densities were chosen (exceptions are the CIFs corresponding to delayed ART in TBM grade I and immediate ART in TBM grade II). Moreover, 14 out of 16 fitted CIFs required a true SNP model with $K \geq 1$ and 3 required $K = 3$.

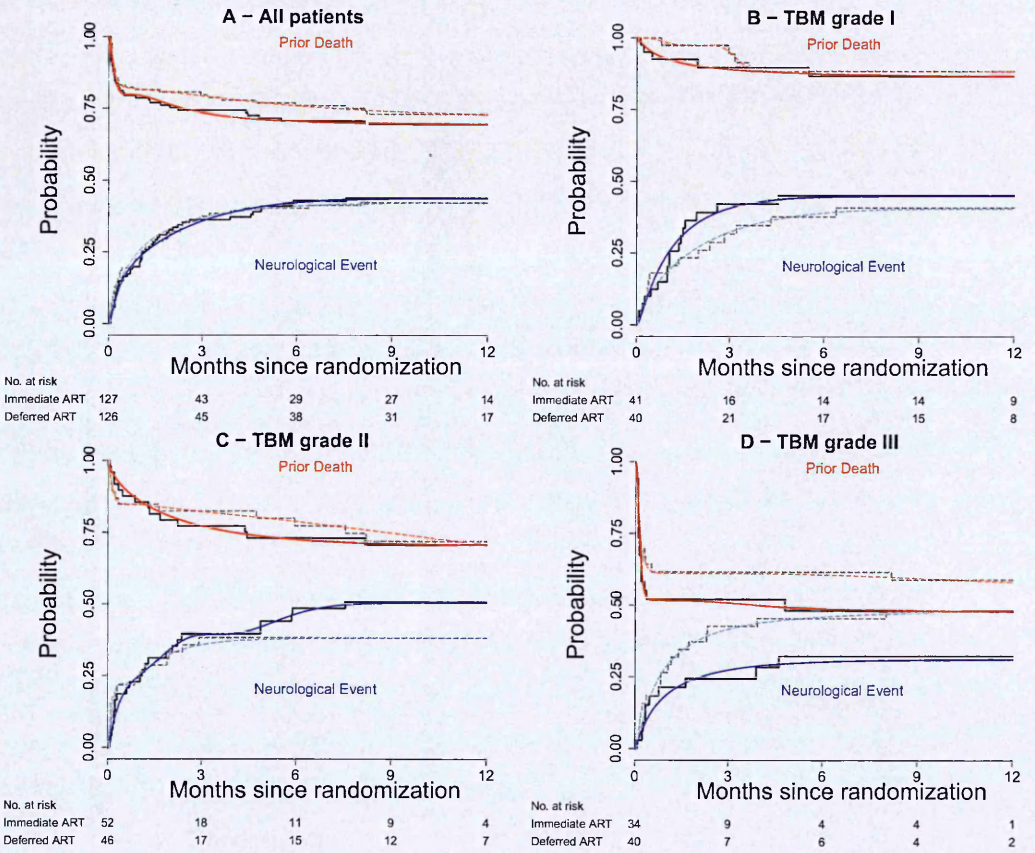
Results from statistical tests to compare the estimated CIFs from SNP models between the two treatment arms based on the IWD statistic (Section 3.5) are given in Table 3.14 with p-values derived from both the delta-method and based on a Monte-Carlo (MC) permutation tests. For these tests, the limit of the integral of the corresponding IWDs was 270 days and the unity weight function ($W \equiv 1$) was used. Table 3.14 shows large p-values for both competing events which is in-line with the fact that observed differences between the two randomized arms appear to be mostly small.

Table 3.14: P-values of IWD-based tests for differences in CIFs of neurological events and prior death between treatments.

Method for p-value Event	Delta-method	MC permutation test*
Neurological event	0.993	0.993
Prior death	0.381	0.400

*: based on 1000 MC samples.

Figure 3.7.1: Cumulative incidence function for the time to the first neurological event (blue) and one minus cumulative incidence function for prior death (red) by treatment arm based on the SNP-AIC method and the nonparametric method.



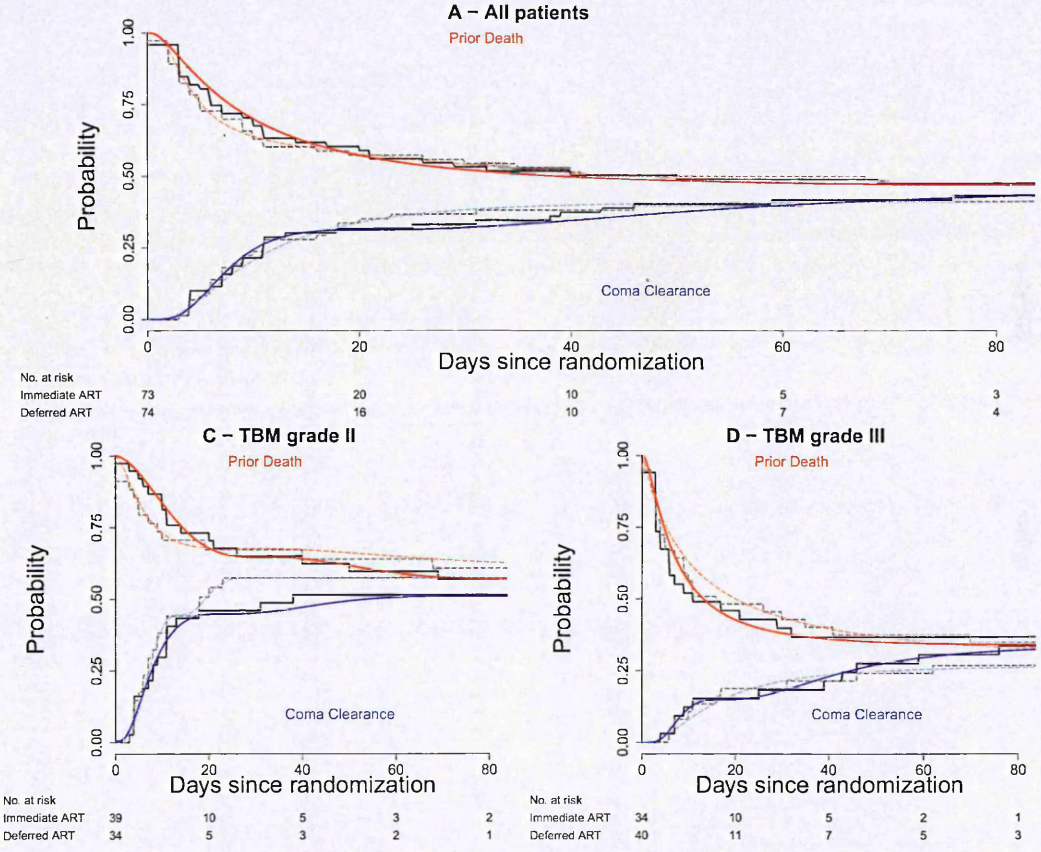
Solid lines correspond to immediate, dashed lines to delayed ART ("placebo"). Black and grey lines correspond to nonparametric estimates.

The outcome time to coma clearance or prior death was evaluated in 147 patients with impaired consciousness (GCS < 15) at enrolment and by definition, this excluded all subjects with TBM grade I and some with TBM grade II. In total 61 coma clearances and 79 prior deaths were observed, and in patients who cleared coma, this took a median (maximum) of 9 (181) days. The corresponding CIF estimates are displayed in Figure 3.7.2. As before my SNP estimated CIFs are also close to the nonparametric estimates and most SNP CIFs have standard normal base densities except for TBM

grade II. I also observed that 3 of the 12 SNP fitted CIFs required $K = 3$ and 7 of them had $K \geq 1$.

Of note, the fact that in general the fitted SNP models chose polynomial degrees > 0 indicates that simple parametric models such as lognormal or Weibull mixture models might fail to capture the precise shapes of some of the CIFs.

Figure 3.7.2: Cumulative incidence function for the time to coma clearance (blue) and one minus cumulative incidence function for prior death (red) by treatment arm based on SNP-AIC method and nonparametric method.



Solid lines correspond to immediate, dashed lines to delayed ART ("placebo"). Black and grey lines correspond to nonparametric estimates.

3.7.2 Combination antifungal therapy for cryptococcal meningitis

In this example, I use a data set from a recent RCT conducted at OUCRU-VN investigating the effect of three different antifungal therapies in HIV-infected patients with cryptococcal meningitis (Day et al. (2013)). The two co-primary study endpoints were mortality during a follow-up period of 14 and 70 days, respectively. The study found that combination antifungal therapy with Flucytosine and Amphotericin B led to a significantly lower mortality at 70 days compared to Amphotericin B monotherapy, the standard of care treatment in Vietnam. In contrast, superiority of a second combination therapy,

Fluconazole and Amphotericin B, over Amphotericin B monotherapy could not be demonstrated.

The rate of decline of quantitative yeast fungal counts in the cerebrospinal fluid is an important marker of the potency of an antifungal therapy and the competing risks outcome considered here is the time from randomization to fungal clearance (beneficial event of interest) or prior death i.e. death without prior fungal clearance (harmful competing event). For this example I consider a follow-up period of 30 days as fungal count measurements were only performed irregularly after that time point. As fungal count measurements were only measured weekly according to protocol, the time to fungal clearance is given as having occurred in the interval from the time of the last positive count (+0.001) until the time of the first count of 0 (i.e. observed fungal clearance). Patients who died without prior fungal clearance were treated as prior deaths, e.g. if a patient died on the second day after randomization, the patient was assumed to have died in the interval [2, 3] days. Patients who neither reached fungal clearance nor died during the 30 days were right-censored at the time of their last positive fungal count. The data set contains data from 263 patients: 155 reached fungal clearance, 64 died without prior clearance, and 44 were censored.

One problem with the endpoint derivation is that as fungal counts are not continuously measured and fungal clearance is not an absorbing state, one cannot be 100% sure that subjects who died without documented fungal clearance truly did not reach (unmeasured) fungal clearance before death. If one wanted to properly account for this, this would require multi-stage modelling of the entire illness-death model which also allows for transitions from fungal clearance to death allowing for unobserved transitions due to interval-censoring. This is beyond the scope of this project. However, it is believed that most likely only very few subjects who were considered as prior deaths had previously reached (unobserved) fungal clearance for the following reasons:

- Amongst the 64 deaths, 40 died in the first 7 days. Fungal clearance before day 7 is very unlikely: amongst 112 fungal counts measured in the original database after enrolment but before day 7, 95.5% (107/112) were positive. Specifically, among counts measured on day 6, 96.6% (28/29) were positive.

- Amongst the 24 subjects who died later than on day 7, linear extrapolation of their last two log-transformed counts was used to assess how likely it was that they were negative on the day of deaths. The prediction only gave likely fungal clearance prior to death in 3 (12.5%) of them.

- The above calculations are conservative as higher fungal counts and lower fungal clearance has been shown to be associated with increased risk of dying (Bicanic et al. (2009) and Day et al. (2013)).

Thus, in the analysed data set I assume all deaths were without prior fungal clearance which is conservative, i.e. it tends to underestimate the effect of antifungals on fungal clearance, a beneficial outcome. In addition, for simplicity, I only consider two treatment arms, namely Amphotericin B monotherapy and Amphotericin B plus Flucytosine, the more potent combination therapy.

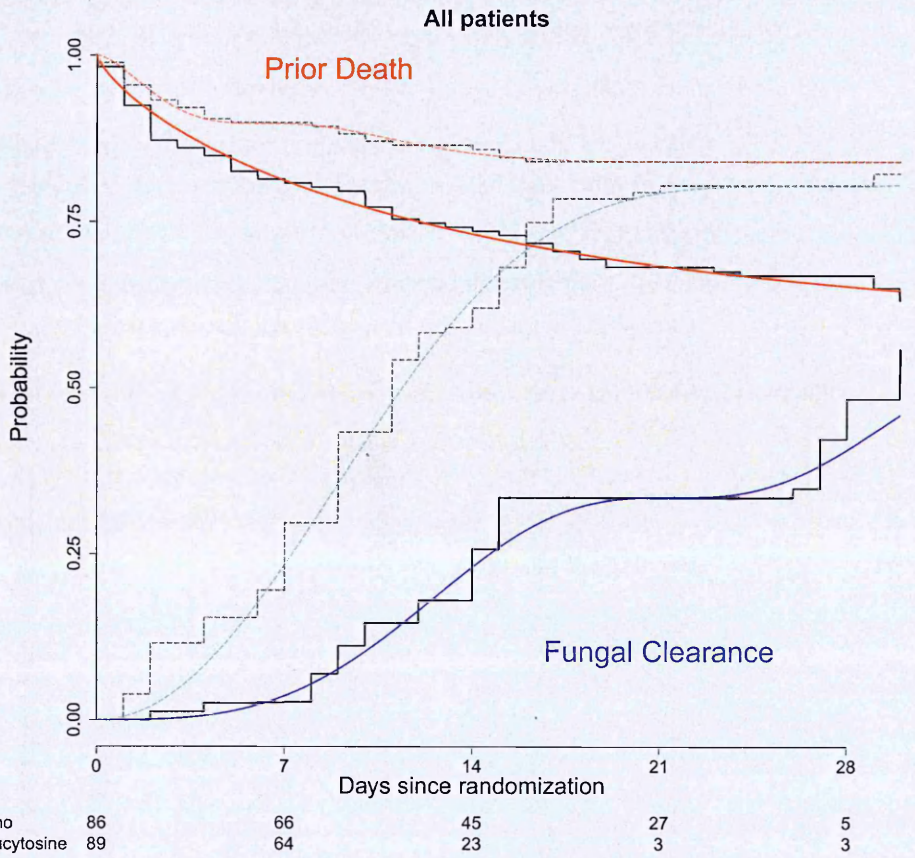
CIF estimates are displayed in Figure 3.7.3 and show both a faster time to clearance and a lower risk of prior death for the combination therapy. This is also reflected by results from IWD-based tests for treatment effect on both competing events as displayed in Table 3.15. In these tests the corresponding IWDs focus on the period from 0 to 30 days with unity weight function. As in the previous example, SNP and nonparametric estimates closely agreed. In all SNP fits to the original data set the respective polynomials had $K \leq 1$.

Table 3.15: P-values of IWD-based tests for differences in CIFs of fungal clearance and prior death between treatments.

Method for p-value Event	Delta-method	MC permutation test*
Fungal clearance	≤ 0.001	≤ 0.001
Prior death	0.020	0.026

*: based on 1000 MC samples.

Figure 3.7.3: Cumulative incidence function for the time to fungal clearance (blue) and one minus cumulative incidence function for prior death (red) by treatment arm based on SNP-AIC method and nonparametric method.



Solid lines correspond to Ampho mono, dashed lines to Ampho + Flucytosine. Black and grey lines correspond to the nonparametric estimates.

3.7.3 Menopause data

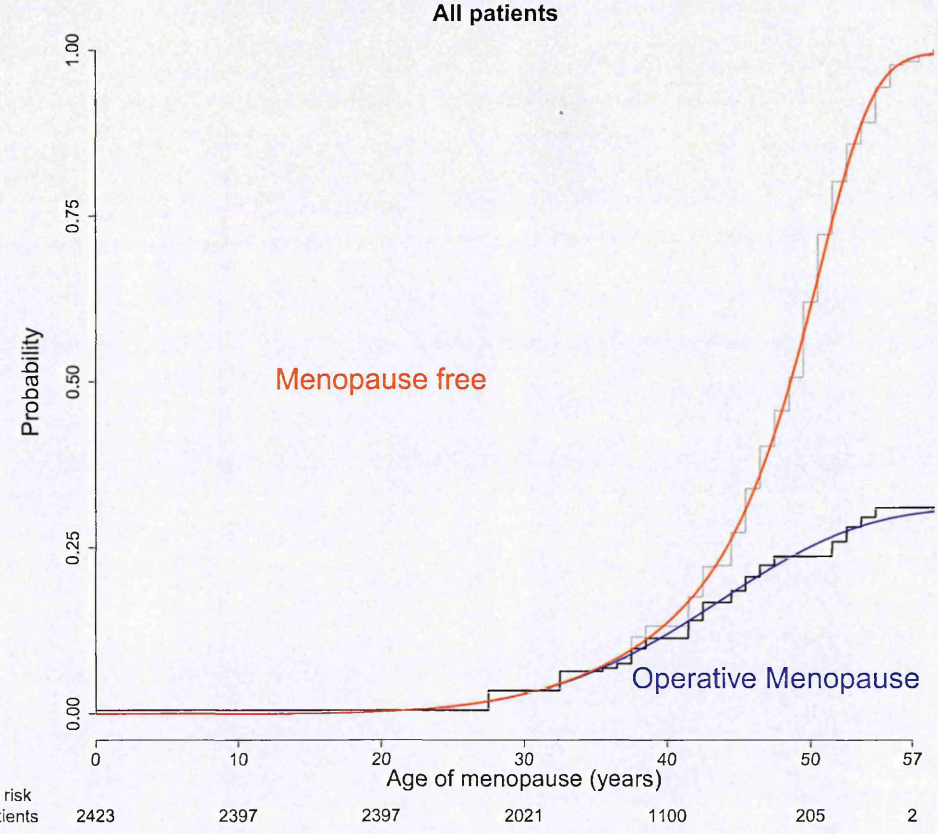
In the previous example, I analysed a data set with interval-censoring. As mentioned in Subsection 1.1.2 one special type of interval-censored data is current status data where the status of the subject is only assessed at one time point. One example of a current status data set with competing risks is the menopause data set described in Krailo & Pike (1983) which is publicly available as data set menopause in the contributed R package `etm`. In brief this data set contains information on menopausal status of 2423 women and the time scale refers to a subject's age. Menopausal status (operative, natural or none) for each subject was assessed only once giving rise to current status data.

As in this setting, it is also of interest to summarize the cumulative proportion of women having any type of menopause, for my SNP estimates, I stacked the two estimated CIFs in the graphical display such that their sum provides an estimate of this quantity. The same was done for the nonparametric

estimates.

Again, from Figure 3.7.4, SNP results and the nonparametric fits agree with each other. For this data set, the best SNP fit is in fact the same as a parametric Weibull model. Despite having almost 70% right-censored cases, i.e. women who did not have menopause at their assessment time point, CIF estimates from both SNP and nonparametric method add up to almost 1 at age 58.5. This can easily be explained by the fact that biologically women are highly unlikely to have menopause (due to any cause) after age 58, so the observed follow-up duration of time to menopause covers a range with very high mass of the distribution of the marginal time-to-event.

Figure 3.7.4: Stacked cumulative incidence function for the time to operative menopause (blue) or natural menopause (red) based on SNP-AIC method.



Area above the red curve corresponds to not having any type of menopause. Black and grey curves correspond to the nonparametric estimates.

3.8 Discussion

In this chapter I have developed a new semi-nonparametric (SNP) model for the CIFs of competing risks data with arbitrary censoring and truncation. I then implemented a corresponding estimation

algorithm in the statistical language R and compared the performance of the proposed estimator to alternative approaches in an extensive simulation study. The algorithm also implemented parametric lognormal and Weibull mixture model for competing risks as a starting point which are also not commonly available in statistical software.

SNP models were designed to combine the advantages of both parametric and nonparametric methods without the respective disadvantages. My simulations study confirmed that this is achieved in many instances and the SNP model frequently outperformed both parametric and nonparametric estimators. This confirms and extends similar results which have been previously obtained in survival analysis (Zhang & Davidian (2008) and Doehler & Davidian (2008)). Another strength of my method is that it is based on parametric submodels and thus easily allows for the inclusion of arbitrary censoring and truncation pattern. I have demonstrated this for interval-censoring. To my knowledge, in this setting there is currently only a single alternative method for CIF estimation implemented in the statistical software R. This alternative is a nonparametric method and the implementation currently provides only estimates without associated confidence intervals. Moreover, my simulation study demonstrates that with interval-censoring, my SNP estimator gave more precise estimates across all simulation scenarios. Even though I only mentioned and implemented a competing risks model for right-censoring, interval-censoring and left-truncation which are most frequently seen in clinical applications, my model can be easily extended to other situations such as left censoring and the general case of interval truncation (see page 436 of Everitt & Skrondal (2007)). In addition, it should be straightforward to extend my model to situations where the event type is partially missing. An additional strength of the method is that it provides smooth estimates which may frequently be more realistic approximation to the truth than a step function and also allow for more realistic simulations based on the fitted model.

The proposed algorithm for SNP CIF estimation in general worked well. Based on the simulation studies, I recommend AIC as the default information criterion for selection of the polynomial degrees because it performed best in situations when I simulated non-SNP scenarios and only lost little in SNP scenarios. As mentioned earlier, my greedy stepwise forward algorithm may not always identify the optimal information criterion but exhaustive search through all models would be much more computationally intensive especially if J, the number of different event types, is large. Of note, the use of information criteria like AIC,BIC and HQC for SNP model selection has been known for choosing too simple models in some applications (Coppejans & Gallant (2002)). Consequently an alternative criterion based on cross validation and mean-squared error was suggested. However, such a method would be practically infeasible in my setting due to a tremendous increase in computing time. The substantial computational power needed for my algorithm also led me to restrict the maximum poly-

nomial degree K_{max} in the simulation study and prevented the exploration of coverage probabilities of bootstrap-based confidence intervals. In practical applications, where only a single model fit is required, I recommend exploring higher polynomial degrees though for all of the real applications that I considered, $K_{max} = 2$ was sufficient. Bootstrap methods might improve upon the occasional observed undercoverage of confidence intervals and indeed, this has been shown to be true in SNP survival models (Zhang & Davidian (2008)).

My model is based on the mixture factorization (3.1.1) of the CIF and used a simple multinomial model for the marginal event probabilities and a SNP survival models for the conditional time-to-event models. As an alternative to my approach, one could explore the performance of other semi-parametric estimators which have been proposed for estimation of the survival function such as the logspline density estimator (Kooperberg & Stone (1992)). However, in the survival context, SNP estimators have been shown to outperform the logspline estimator (Doehler & Davidian (2008)) and I did not pursue this further.

A limitation of the SNP approach is that despite its strong performance in simulation studies asymptotic properties such as consistency and asymptotic normality have not yet been established in the survival or the competing risks setting. I conjecture that it may be possible to extend the existing consistency proof for SNP density estimation (Gallant & Nychka (1987)). In the competing risks setting, one challenge is that the marginal probabilities of the different event types which are required in the mixture factorization are non-identifiable in the presence of a limited observed follow-up duration. Interestingly, this did not seem to deteriorate the performance of the SNP estimator within the observed follow-up period in the simulation studies despite heavy simulated censoring, but it could complicate a consistency proof for the CIF up to the maximum follow-up. A plausible direction to tackle this issue could be found in Maller & Zhou (2002), who provided the conditions for achieving consistent estimates of parameters in parametric models based on mixture factorization. These include "sufficient follow-up" which is usually satisfied when right-censoring is not too heavy. Additionally, my simulation study demonstrated that confidence intervals assuming asymptotic normality of the estimator often performed well and this may indicate that the resulting estimator is indeed asymptotically normally distributed. However, a formal proof of this would be extremely challenging.

Finally the current estimator does not include covariate information and I shall discuss SNP competing risks regression modelling in the next chapter.

Chapter 4

CIF-based regression method using SNP densities

Regression models for competing risks which model the cumulative incidence function (CIF) are appropriate for clinical applications where the main interest is in the absolute risks of events occurring over time. As discussed in Chapter 1, this is particularly relevant for prognostic research and medical decision making. In this chapter I shall discuss extension of the CIF-estimation approach introduced in the previous chapter to regression modelling.

4.1 Model formulation

Following the notations in Chapter 3, a direct extension of the CIF estimation approach proposed in Section 3.1 to regression modelling is achieved by including covariates into Equations (3.1.2) and (3.1.3). For event type j where $j = 1, ..., J$ this results in:

$$
\begin{aligned}
P(T \leq t, D = j; \mathbf{Y}_j, \mathbf{X}) &= P(T \leq t \mid D = j; \mathbf{Y}_j)\, P(D = j; \mathbf{X}) \\
\log(T \mid D = j; \mathbf{Y}_j) &= \mathbf{Y}_j^T \beta_j + \log\{T_{0j}\} = \mathbf{Y}_j^T \beta_j + \mu_j + \sigma_j Z_j \text{ for } j = 1, ..., J \\
P(D = j; \mathbf{X}) &= \frac{\exp(\mathbf{X}^T \gamma_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{X}^T \gamma_j)} \text{ for } j = 1, .., J-1 \text{ and} \\
P(D = J; \mathbf{X}) &= \frac{1}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{X}^T \gamma_j)}
\end{aligned}
\tag{4.1.1}
$$

where $\mathbf{X} = (1, X_1, \ldots, X_p)$ and $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jq_j})$ are covariate vectors for the marginal $(P(D = j; \mathbf{X}))$ and the conditional $(P(T \leq t \mid D = j; \mathbf{Y}_j))$ components, respectively. The corresponding vectors of regression coefficients are $\gamma_j = (\gamma_{j0}, \gamma_{j1}, \ldots, \gamma_{jp})$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jq_j})$.

In Model (4.1.1), the conditional time to event distributions $T \mid D = j$ are modelled with standard

AFT models. As in Chapter 3, I assume that Z_j either has a SNP-distribution with a standard normal base density or $\exp(Z_j)$ has a SNP-distribution with an exponential base density with rate 1. In this model, T_{0j} can be interpreted as the covariates-free "baseline" conditional survival time. For modelling the marginal probabilities, $P(D = j; \mathbf{X})$, I use a standard multinomial logistic model. For this model, the choice of the baseline category J does not affect the model fit but changes the interpretation of the regression parameters γ_j.

While the multinomial logistic model by definition includes only a single set of covariates $\mathbf{X}$, the conditional time-to-event distributions $T|D = j,\ j = 1, \ldots, J$ could in principle depend on separate sets of covariates $\mathbf{Y}_j$. However, in practical applications, it will frequently be most meaningful to use the same or similar sets of covariates for the logistic model and all conditional time-to-event distributions. Note that with a slight abuse of notation I use γ_j to denote a vector of regression coefficients in this chapter while γ_j in Equation (3.1.3) of Chapter 3 is a scalar. In addition, note that although X contains a constant to model an intercept, the $\mathbf{Y}_j$ do not. This is to avoid non-identifiability of the AFT model which already absorbs the intercept into μ_j.

For survival analysis based on SNP densities, Zhang & Davidian (2008) suggested three alternative models: the AFT chosen here, the proportional-hazards model and the proportional-odds model. In principle, all three of them could be used to model the conditional distribution of $T \mid D = j$ in competing risks. I chose the AFT model as the basis for my own models for the following reasons: First the use of the AFT model provides a direct way to assess covariate effects on time rather than a derived quantity such as the hazard function which facilitates interpretation. Second, only the AFT model is "closed" under a family of SNP distribution with any pre-specified base density. This means for competing risk j that both the baseline conditional log survival time, $\log(T_{0j}) = \mu_j + \sigma_j Z_j$, and all covariate-dependent conditional log survival time, $\log(T \mid D = j; \mathbf{Y}_j)$, in Model (4.1.1) follow essentially the same distribution (i.e. identical base density and SNP-polynomial) other than the intercept term. This would not be the case for e.g. the proportional hazards model where only T_{0j} but not covariate-dependent survival times would follow a "simple" SNP model and, hence, the model would depend on the scaling of covariates which is undesirable.

4.2 Parameter estimation and ad hoc statistical inference

The likelihood construction for the regression model (4.1.1) is essentially identical to the likelihood construction required for estimation of the CIF detailed in Section 3.2, except for the fact that the likelihood now also depends on additional covariates. Likewise, the same estimation procedure as detailed in Section 3.3 can be used with the additional need for acquiring starting values for the

regression coefficients β_j and γ_j ($j = 1, \ldots, J$). This is done as follows.

Starting values for the parametric conditional survival models, i.e. μ_j, σ_j and β_j are obtained by fitting lognormal or Weibull AFT regression models as described in Subsection 3.3.1. As described in that subsection, subjects experiencing events other than j were excluded from the AFT model for event type j and right-censored observations were weighted according to a crude estimate of their probability of ultimately experiencing event type j conditional on being event-free at their censored times. For simplicity, I used the same (covariate-independent) estimate of this probability as for the case without covariates but included the covariates X_j into the AFT to obtain starting values for β_j. Elements of γ_j are initialized by using the suboptimization described in Subsection 3.3.1, for which starting values of the intercepts γ_{j0} are computed as previously described in Equation (3.3.1) while the non-intercept elements in γ_j are set to 0s. This initialization of the intercept implicitly assumes that a covariate value of 0 corresponds to an "average" observation and for this reason, I recommend centering the covariates prior to fitting the model for numerical reasons. This does not affect the estimates of the regression parameters except for γ_{j0} and μ_j which can easily be transformed back to the scale of the original covariates if needed.

The whole estimation procedure is then carried out following the same step-wise forward algorithm as described in Subsection 3.3. For each forward step when the SNP-polynomial degree K_j of the conditional survival distribution of event type j is increased, multiple starting values for the spherical coordinates from a grid are chosen as previously described. The corresponding initial values for μ_j and σ_j are computed similar to Subsection 3.3 with a slight change of notation that $T \mid D = j$ in Equation (3.3.2) is now the baseline $T_{0j} \mid D = j$ as introduced in Model (4.1.1). Starting values for β_j and γ_j are the best fits from the previous step.

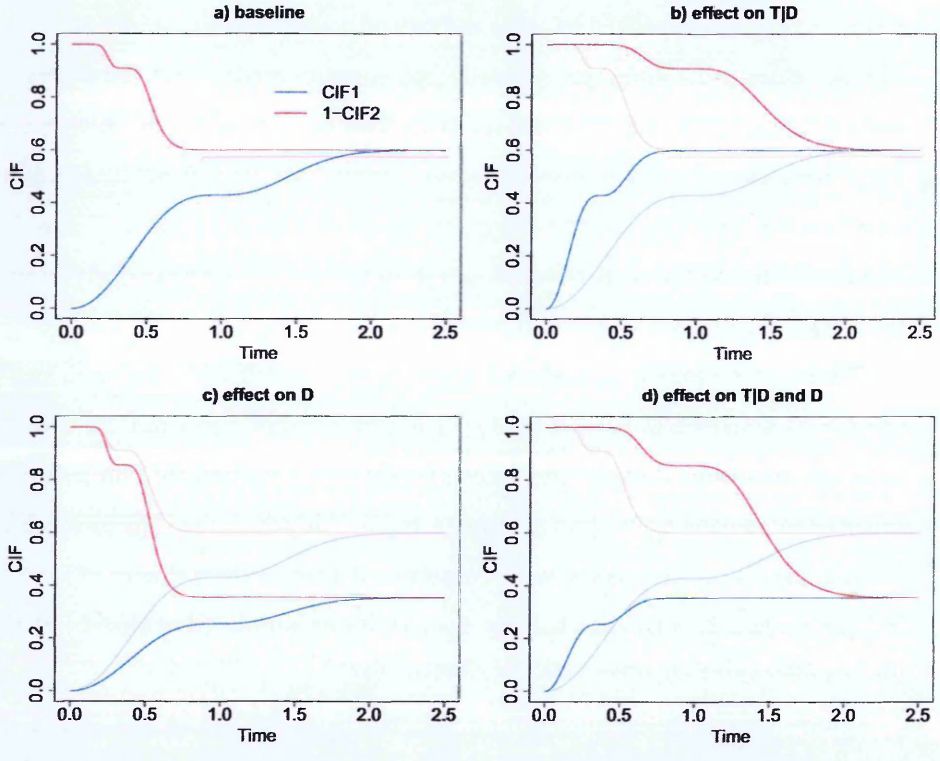
Approximate statistical inference for the SNP regression model (4.1.1) can also be performed using asymptotic theory for maximum likelihood estimation and ignoring the adaptive choice of the polynomial degree as described in Section 3.4. For testing whether a covariate affects the entire competing risks process at all, one can use a Wald-type test which simultaneously tests whether all regression coefficients (from the multinomial component model and all AFT sub-models) associated with that covariate can jointly be 0.

4.3 Model illustration and interpretation of parameters

An illustration of how a single binary covariate can affect the resulting CIFs in Model (4.1.1) is displayed in Figure 4.3.1. In Figure 4.3.1 a), the baseline CIFs for two competing events ($J = 2$) are shown. First, assume that the binary covariate does not alter the marginal event probabilities but ac-

celerates the conditional time to the first event type T_{01} by a multiplicative factor of e^{-1}(i.e. $\beta_1 = -1$) while decelerating the conditional time to the second event T_{02} by a multiplicative factor of e (i.e. $\beta_2 = 1$). The resulting CIFs are shown in Figure 4.3.1 b). Second, assume that the binary covariate is associated with an odds ratio of $exp(-1) \approx 0.37$ for the probability of the first event but does not affect the conditional time-to-event distributions. The resulting CIFs are shown in Figure 4.3.1 c). Third, if both covariate effects are combined, the resulting CIFs are shown in Figure 4.3.1 d).

Figure 4.3.1: Illustration for different covariate effects on SNP model.



a) Baseline CIFs (pale curves in other panels), b) CIFs for a binary covariate affecting only $T \mid D = j$, c) CIFs for a binary covariate affecting only D, c) CIFs for a binary covariate affecting both $T \mid D = j$ and D.

The component models (AFT and multinomial logistic) are well-known statistical models that are frequently used in practice and their interpretation is relatively straightforward though the interpretation of regression coefficients for the multinomial logistic model with more than two event types ($J > 2$) can be somewhat involved. As illustrated in Figure 4.3.1 c), the proposed model is flexible and allows to model quite general covariate effects on the CIF (including crossing CIFs) if covariates are allowed to affect both the marginal and conditional components. However, this also complicates interpretation of the model because it is difficult to assess how the combined effect of a covariate on different component models jointly influences the resulting CIFs. Hence, in clinical applications, it will sometimes be beneficial to include covariates for selected component models only to facilitate

interpretation if such a simplified model does not markedly deteriorate the model fit.

If covariates are only allowed to the multinomial logistic model i.e. the marginal component, then

$$\log CIF_j(t) = \mathbf{X}^T\gamma_1 - \log\left(1 + \sum_{k=1}^{J-1} \exp\left(\mathbf{X}^T\gamma_k\right)\right) + \log P\left(T \leq t \mid D = 1\right) \tag{4.3.1}$$

for $j = 1, \ldots, J-1$, and a similar model holds for $j = J$. This implies that the covariates have a multiplicative (and hence time-homogeneous) effect on the resulting CIFs. When $J = 2$, γ_1 is simply the covariate vector containing the log-odds ratios corresponding to the marginal probability of event type 1 occurring. Hence a positive log-odds ratio implies that an increase in the corresponding covariate is associated with an increase in CIF_1 at all time-points and a decrease in the CIF_2 at all time-points. This simple qualitative interpretation of covariate effects on the CIF scale is similar to the Fine and Gray model (Fine & Gray (1999)) introduced in Section 1.5.2 except that it models covariate effects on different CIFs simultaneously and in a "consistent" way. As an example, this implies that their sum at any time point cannot be greater than 1 for any covariate combination which is not necessarily the case if CIF estimates are based on different Fine and Gray models for different event types. Indeed, as mentioned before in 1.5.2, in most settings it is mathematically impossible that exact Fine and Gray models hold simultaneously for all event types.

Interpreting covariate effects on the conditional component has been criticised because the corresponding survival time is conditional on the future event type (Andersen & Keiding (2012)). Nevertheless, one can think of situations where such a parametrization could be useful such as an intervention that is expected to delay the conditional time to some detrimental outcome but not to change the ultimate probability of its occurrence. However, even if the intervention is randomized, causal interpretation of the resulting effects on the conditional components is delicate as it conditions on events occurring post randomization.

4.4 Limitation of the proposed model for studies with limited follow-up and an alternative model

As shown in Section 4.3, the proposed regression model (4.1.1) allows to model covariate effects on the CIF in a flexible way but it has an additional important limitation which has not been mentioned so far: The multinomial logistic sub-model describes the distribution of the marginal event probabilities, $P(D = j)$, but these event probabilities might be poorly identified based on data from studies with a limited duration of follow-up and heavy right-censoring.

As a simple example with only two event types ($J = 2$) and no covariates, assume that the nonpara-

metric CIF estimates at the maximum observed follow-up duration $t_{\max}$ in a study are $\widehat{CIF}_1(t_{\max}) = 0.15$ and $\widehat{CIF}_2(t_{\max}) = 0.25$. Even if these estimates were exact, all that the data would imply regarding $P(D = 1)$ is that it must lie between 0.15 and 0.75, i.e. a non-informative bound. The implication for my model is that in this case, the model parameters would be nearly non-identifiable and obtained model-estimate would be dubious as they would heavily depend on the constraints implied by the (semi-)parametric model formulation rather than the actual data.

Interestingly, the simulation in Chapter 3 indicates that the CIF can be quite accurately estimated at time-points t when there are still a substantial number of patients under follow-up even if the underlying data is heavily censored. Thus, the identifiability issues of the underlying model did not appear to corrupt estimation of the CIF, i.e. the probabilities $P(T \leq t, D = j)$, over the observed time range of follow-up. This may indicate that risk predictions from my regression models could also be relatively unaffected by a limited follow-up duration if predictions are restricted to the observed time range and I will further investigate this in the simulation study in Section 4.5. Moreover, if covariates are only allowed to affect the marginal probabilities, this induces time-homogeneity of covariate effects as discussed in Section 4.3, and under this constraint, the estimation of covariate effects is expected to be less affected by identifiability problems.

Nevertheless, the raised limitation indicates that my regression model is most applicable in settings with substantial follow-up relative to the timing of events, i.e. if the marginal probabilities are relatively well-identified based on the observed data. Otherwise, covariate effects should only be interpreted with extreme caution.

One alternative approach that avoids identifiability issues involving $P(D = j)$ and still adheres to the mixture factorization is to model the event status D at a time-point where it is expected to be well-identified by the data explicitly rather than at time infinity. This might also be of interest when accurate risk predictions at a specific time point are desired, e.g. 10-year risk predictions for the occurrence of coronary heart disease which form the basis for treatment guidelines (Wolbers et al. (2009)). Specifically, choose a time point t_m which I assume to be later than or equal to the maximum observed event or follow-up time $t_{\max}$ and define a new random variable

$$D_m = \begin{cases} D & \text{, if } T < t_m \\ 0 & \text{, otherwise} \end{cases}$$

Obviously $D_m = D$ for $t_m = \infty$ due to the basic assumption of most competing risks models that all subjects are assumed to have an event eventually. Under this new setting, the mixture factorization in

(3.1.1) becomes

$$\begin{aligned} P(T \leq t, D = j) &= \frac{P(T \leq t, D = j)}{P(T < t_m, D = j)} P(T < t_m, D = j) \\ &= \frac{P(T \leq t \mid D = j) P(D = j)}{P(T < t_m \mid D = j) P(D = j)} P(D_m = j) \\ &= \frac{P(T \leq t \mid D = j)}{P(T < t_m \mid D = j)} P(D_m = j) \end{aligned} \tag{4.4.1}$$

In the second equality, $P(T < t_m, D = j) = P(D_m = j)$ by the definition of D_m. This mixture factorization is only sensible for time points t with $t < t_m$ but as I have assumed that t_m is larger than all observed event or censoring times, it can be used as the basis of a likelihood construction which can essentially proceed as described in Section 3.2 while replacing $P(D = j)$ with $\frac{P(D_m=j)}{P(T<t_m|D=j)}$.

For this alternative model, one can still use a multinomial model for D_m but allow for an extra category to account for the possibility of $D_m = 0$. Moreover, the conditional components can be modelled based on AFT models and SNP densities as before.

One problem of the above model is that even though I directly model the event probabilities at time t_m, the time-to-event random variable T still has support on $[0, \infty)$ even though in reality, this distribution is also non-identifiable beyond t_m. Indeed, when I implemented the model, the estimation procedure either crashed or provided highly variable CIF-estimates even for parametric scenarios without covariates. I suspect that this is an issue of the numerical non-identifiability of $P(T \leq t \mid D = j)$ due to the presence of $\frac{P(T\leq t|D=j)}{P(T<t_m|D=j)}$ in the likelihood function. Specifically, two different specifications for $P(T \leq t \mid D = j)$ can result in numerically very similar likelihood contributions of $\frac{P(T\leq t|D=j)}{P(T<t_m|D=j)}$ for $t \leq t_m$. Indeed, if the two specifications lead to proportional but non-identical conditional cumulative distributions up to time t_m, their likelihood is identical. Due to these shortcomings and the failure of my numerical algorithm in this setting, I did not pursue this approach further. Yet an alternative approach could be to use distributions with support on $[0, t_m]$ as the basis for modelling the CIF up to time t_m. In fact, SNP models for base densities with bounded support were introduced by Kim (2007). However, exploring such an approach in detail is outside the scope of my thesis.

4.5 Simulation study

In this section I report a simulation study to assess the performance of the SNP regression model defined in Equation (4.1.1) under various settings. One aim of this simulation study is to assess the potential bias and precision of parameter estimates. The second aim is to evaluate the precision of covariate-dependent CIF-estimates of my model if data are either simulated according to the mixture model (4.1.1) or alternative popular regression models for competing risks.

For all scenarios, I considered two competing risks ($J = 2$) and two independent covariates X_1 and X_2. X_1 was simulated according to a Bernoulli distribution with $p = 0.5$, which could for example represent a random treatment assignment in an RCT. X_2 was simulated according to a normal distribution with mean zero and a standard deviation of 0.5. Uncensored competing risks data were then simulated either according to the mixture factorization (4.1.1) or alternative popular regression models for competing risks as detailed in Subsections 4.5.1-4.5.3, and additional censoring was simulated as described in Subsection 4.5.4.

For each scenario I varied the sample size between 100 and 500. Results for each of the 26 simulation settings (see below) are based on 200 simulated data sets.

4.5.1 Mixture factorization scenarios

The first set of scenarios is based on the mixture factorization in Equation (4.1.1) for which the conditional times to event type j are specified as:

$$\log(T \mid D = j; X_1, X_2) = X_1\beta_{j1}^M + X_2\beta_{j2}^M + \log(T_0 \mid D = j),$$

where the "baseline" distributions of $T_0 \mid D = j$ were chosen as displayed in Table 4.1. The marginal event probabilities are determined as:

$$P(D = 1 \mid X_1, X_2) = \frac{\exp(\gamma_0 + X_1\gamma_1 + X_2\gamma_2)}{1 + \exp(\gamma_0 + X_1\gamma_1 + X_2\gamma_2)}$$

Values of γ_0, γ_1, γ_2 and β_{11}^M, β_{12}^M, β_{21}^M, β_{22}^M are given in the first row of Table 4.1 and were fixed for all distribution configurations of $T_0 \mid D = j$. For convenience, in all subsequent discussions β_{11}^M, β_{12}^M, β_{21}^M and β_{22}^M will be referred to as AFT (regression) parameters whereas γ_0, γ_1 and γ_2 will be called multinomial logistic (regression) parameters. Of note, γ_0 is set to $\log\frac{P_1}{1-P_1}$ with $P_1 = P(D = 1 \mid X_1 = X_2 = 0) = 33\%$ which implies that the marginal probability $P(D = 1)$ is approximately 29%.

Figures 4.5.1, 4.5.2 and 4.5.3 show the resulting CIFs for all 3 mixture factorization scenarios and selected covariates values. Simulated data from these mixture factorization scenarios were subject to both right- or interval-censoring as will be discussed in Subsection 4.5.4.

Table 4.1: Mixture factorization based scenarios.

	Regression parameters (identical for all scenarios)				
	γ_0	γ_1	γ_2	$(\beta_{11}^M,\ \beta_{12}^M)$	$(\beta_{21}^M,\ \beta_{22}^M)$
	-0.71	-0.5	1	$(-2, 1)$	$(-1, 0)$
Scenario names	Baseline distributions				
	$f_{T_0\|D=1}$			$f_{T_0\|D=2}$	
2×Weibull	$W(1, \exp(-1))$			$W(5, \exp(-0.25))$	
2×SNPN	$LN(-1, 0.9^2)$			$SNPN(0.1, 0.8, \frac{\pi}{5})$	
2×logmixturenorm	$0.3LN(1.2, 0.9^2) + 0.7LN(0, 0.6^2)$			$0.5LN(0, 0.1^2) + 0.5LN(1, 0.2^2)$	

Note: $f_{T_0|D=j}$ is the conditional density of the baseline time to each event type j. $W(shape,\ scale)$ means the density of a Weibull distribution with a specific shape and scale. $SNPN(\mu, \sigma, \phi)$ is the density of a random variable T whose $\log T = \mu + \sigma Z$; where Z has a SNP distribution with standard normal base density and spherical coordinates ϕ as described in Zhang & Davidian (2008). $LN(\mu, \sigma^2)$ refers to the density of a lognormal distribution with parameters μ and σ.

Figure 4.5.1: True CIFs for the 2×Weibull scenario.

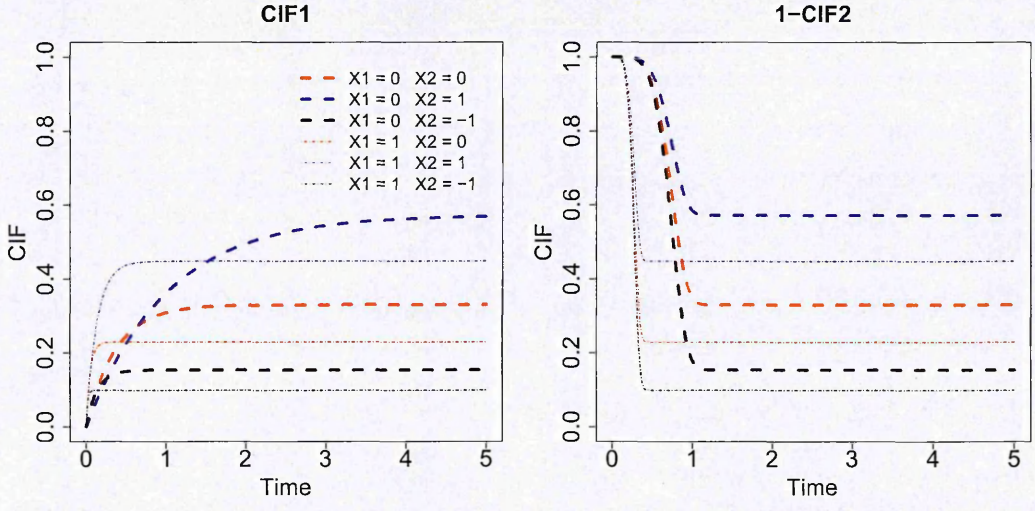


Figure 4.5.2: True CIFs in 2×SNP stdnorm scenario.

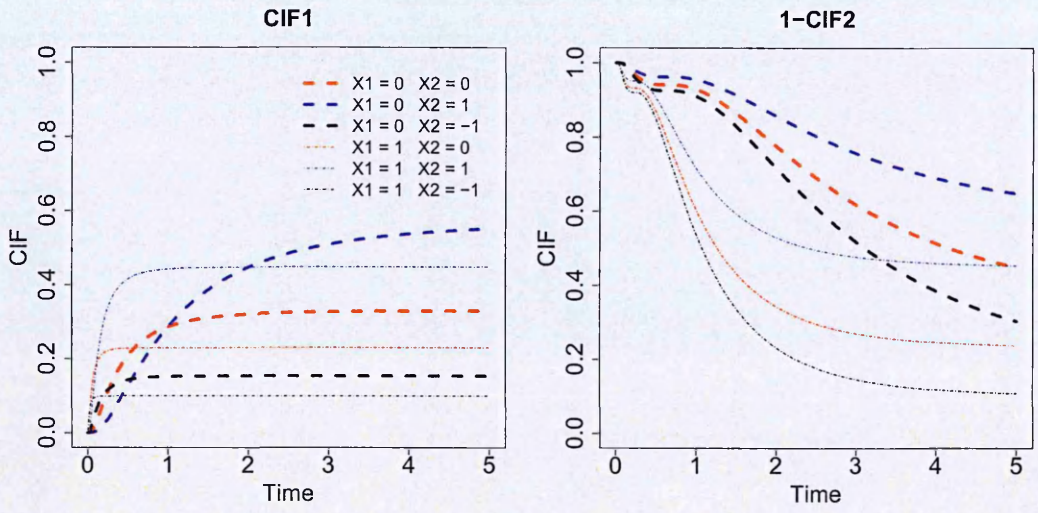
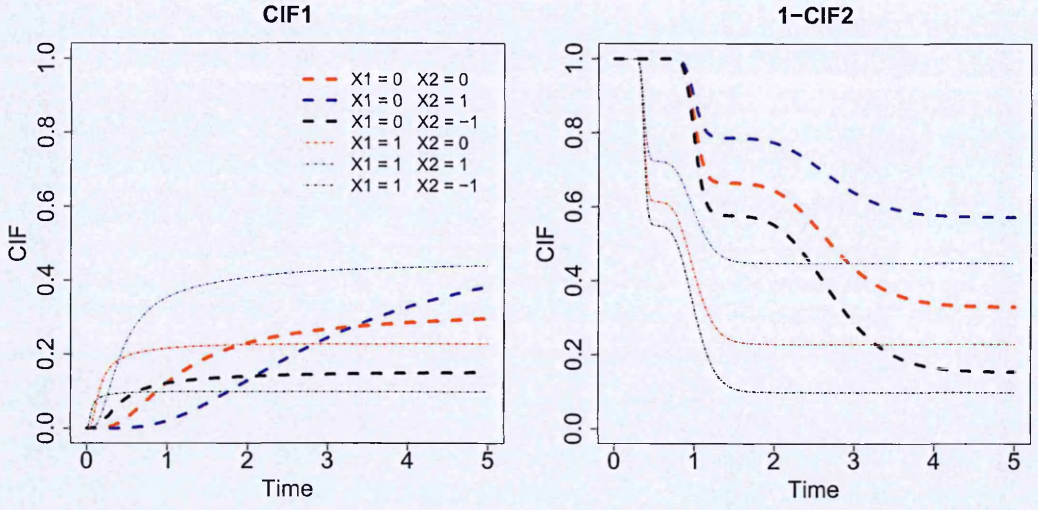


Figure 4.5.3: True CIFs in 2×logmixturenormal scenario.



4.5.2 Fine and Gray's scenario

The second type of scenarios follows the proposed model and simulation in Fine & Gray (1999). Specifically the first CIF follows a Fine and Gray model and is specified as:

$$CIF_1(t; X_1, X_2) = 1 - [1 - p\{1 - \exp(-t)\}]^{\exp\left(X_1\beta_{11}^{FG} + X_2\beta_{12}^{FG}\right)}$$

with $\beta_{11}^{FG} = -1$, $\beta_{12}^{FG} = 1$ and $p = P(D = 1; X_1 = 0, X_2 = 0) = 33\%$. When there are no covariate effects, $CIF_1(t; X_1 = 0, X_2 = 0) = p\{1 - \exp(-t)\}$ is simply p times the cumulative distribution

function of a unit exponential distribution evaluated at t.

The simulation then proceeds as follows: First the event type for each subject is determined via a Bernoulli trial with

$$P(D = 1; X_1, X_2) = CIF_1(\infty; X_1, X_2) = 1 - (1 - p)^{\exp\left(X_1\beta_{11}^{FG} + X_2\beta_{12}^{FG}\right)}$$

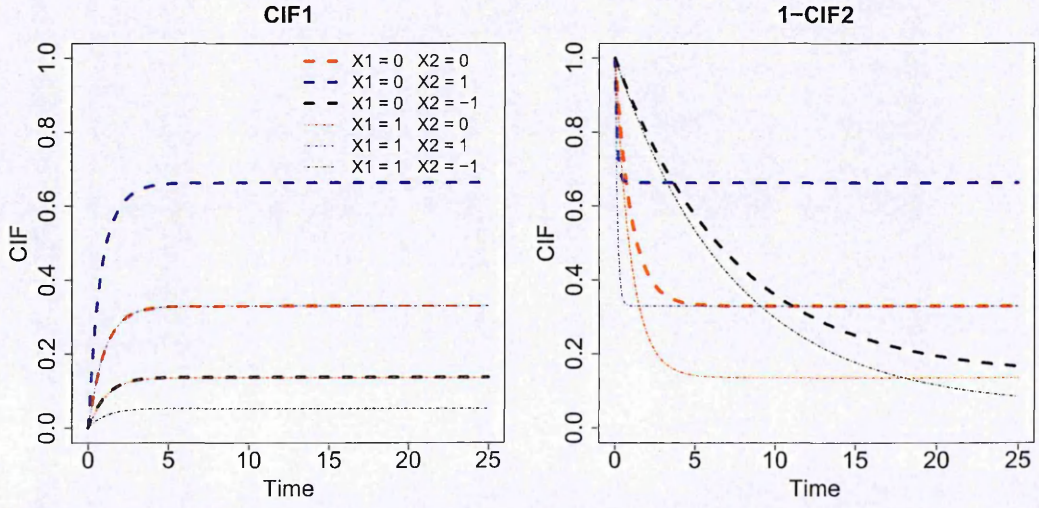
If event type 1 is chosen then the covariate dependent time to event is simulated as

$$T = CIF_1^{-1}(U_1; X_1, X_2)$$

where $CIF_1^{-1}(\cdot; X_1, X_2)$ is the inverse of $CIF_1(\cdot; X_1, X_2)$, and U_1 is a random variable with a uniform distribution on the interval $\left[0,\ 1 - (1 - p)^{\exp\left(X_1\beta_{11}^{FG} + X_2\beta_{12}^{FG}\right)}\right]$. On the other hand if event type 2 is chosen, the conditional distribution of $T \mid D = 2$ was simulated according to an exponential distribution with rate $\exp\left(X_1\beta_{21}^{FG} + X_2\beta_{22}^{FG}\right)$, with $\beta_{21}^{FG} = 0$ and $\beta_{22}^{FG} = 2$. Note that as in the original simulation in Fine & Gray (1999), the CIF for event type 2 does not follow a Fine and Gray model because, as mentioned before, in most settings Fine and Gray models cannot simultaneously be true for all event types.

For the above set-up the marginal probability $P(D = 1)$ is about 25%. CIFs with different covariate values are plotted in Figure 4.5.4. Data simulated from this scenario are subject to only right-censoring.

Figure 4.5.4: CIFs Fine and Gray.



4.5.3 Cause-specific hazards scenarios

The last type of scenario is based on the cause-specific hazards specification as described in Beyersmann et al. (2009). Accordingly the CSH of each competing risk are specified as

$$\lambda_1(t) = \lambda_{01}(t)\exp\left(X_1\beta_{11}^{CSH} + X_2\beta_{12}^{CSH}\right) \text{ and } \lambda_2(t) = \lambda_{02}(t)\exp\left(X_1\beta_{21}^{CSH} + X_2\beta_{22}^{CSH}\right)$$

where $\lambda_{01}(t) = \frac{1}{t+1}$, $\lambda_{02}(t) = 2t$ and $\beta_{11}^{CSH} = -1, \beta_{12}^{CSH} = 1, \beta_{21}^{CSH} = 0$ and $\beta_{22}^{CSH} = 2$. Given these, the total survival time T for each subject is simulated by using the inverse transform sampling method based on the relation:

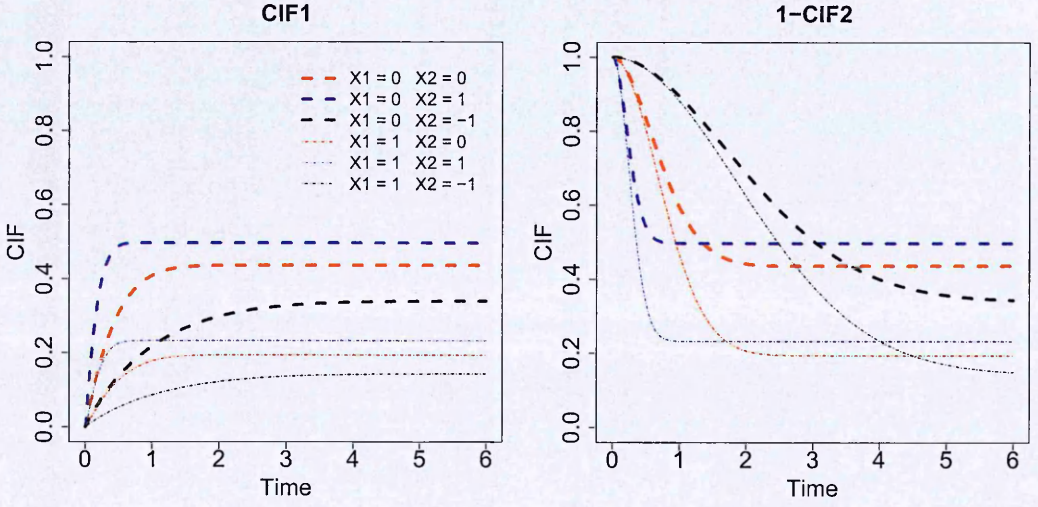
$$P(T \geq t) = \exp\left(-\int_0^t [\lambda_1(u) + \lambda_2(u)]\, du\right) = \exp\left(-\left[\log(t+1)\, e^{X_1\beta_{11}^{CSH} + X_2\beta_{12}^{CSH}} + t^2 e^{X_1\beta_{21}^{CSH} + X_2\beta_{22}^{CSH}}\right.\right.$$

Next for a simulated T a Bernoulli trial is conducted to decide with probability $\lambda_1(T)/\left(\lambda_1(T) + \lambda_2(T)\right)$ if the conditional event type $D \mid T$ equals 1 or not based on

$$\begin{aligned} P\left(D = 1 \mid T \in [t, t+dt), T \geq t\right) &= \frac{P(T \in [t, t+dt), D = 1 \mid T \geq t)}{P(T \in [t, t+dt) \mid T \geq t)} \\ &= \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)} \end{aligned} \tag{4.5.1}$$

The above setting leads to a marginal probability $P(D = 1)$ of approximately 31%. Figure 4.5.5 displays the CIFs for different covariate values in this CSH scenario. As in Fine and Gray scenario, I consider only right-censoring.

Figure 4.5.5: CIFs for the cause-specific hazards scenario.



4.5.4 Simulation of censoring

Two-different types of right censoring (mild and heavy) and one type of interval censoring were simulated for all scenarios. In accordance with Subsection 3.6.1, right-censored data (T', D') was simulated based on the uncensored data (T, D) and censoring time C as follows:

$$T' = \min\{T, C\}, \; D' = \mathbf{I}(T \leq C) \times D$$

As before, C was simulated as $C = \min\{t_m, C_E\}$ where t_m represents the maximum follow-up duration which was set to the 95% or 75% quantiles of the marginal survival time T, respectively. For each of these values of t_m, C_E was generated from an exponential distribution with rate chosen to achieve, respectively, a 10% (mild) or 50% (heavy) overall right-censoring probability.

Interval censoring was simulated by first right-censoring observations at t_m which was set to the 90% quantile of T (implying a 10% right-censoring probability). Then $(0, t_m)$ was divided into 12 equally spaced intervals, and simulation of interval-censoring for each subject proceeded according to Subsection 3.6.2.

4.5.5 Competing approaches

In all scenarios, I used my SNP method with a maximal polynomial degree of $K_{max} = 2$ and the choice of the most appropriate base density and polynomial degrees was based on AIC. For comparison purposes, I used parametric models based on the mixture factorization with lognormal or Weibull models for the conditional AFT models, which are derived from the SNP model by setting $K_{max} = 0$ as mentioned in Subsection 3.6.1. Additionally, for all settings without interval-censoring, semi-parametric alternative models were the Fine and Gray model (or, more precisely, two separate Fine and Gray model for each event type) and the Cox proportional CSH model.

4.5.6 Assessment methods

For the mixture factorization scenarios in Table 4.1, I summarized the mean and standard deviation of the obtained regression coefficient estimates, as well as the mean squared error across simulation runs. To assess reliability of standard MLE-based inference, I also determined the coverage of asymptotic 95% confidence intervals.

For the Fine and Gray and CSH scenarios, looking at estimates of regression parameters from the SNP model is no longer meaningful. Instead, I compared the precision of resulting CIF estimates to the truth which is done via a modification of the AISE in Equation (3.6.1), the so-called mean of

average integrated square error (MAISE) which was also reported for mixture factorization scenarios. In particular for a simulation setting with N_S data sets, the MAISE of event type j is

$$MAISE_j = \frac{1}{N_S}\sum_{i=1}^{N_S}\left[\int_{x_1,x_2}\int_{t=0}^{t_m}\left\{\widehat{CIF}_{ji}(t;x_1,x_2) - CIF_j(t;x_1,x_2)\right\}^2 dtdF(x_1,x_2)\right] \quad (4.5.2)$$

As X_1 is a Bernoulli random variable with probability 0.5, Equation (4.5.2) reduces to

$$\frac{1}{2N_S}\sum_{i=1}^{N_S}\left[\sum_{x_1\in\{0,1\}}\int_{x_2}\int_{t=0}^{t_m}\left\{\widehat{CIF}_{ji}(t;x_1,x_2) - CIF_j(t;x_1,x_2)\right\}^2 dtdF(x_2)\right] \quad (4.5.3)$$

However, exact evaluation of the MAISE is computationally intensive. As a pragmatic solution, the following approximation is used. First, the integrated square difference i.e.

$$\int_{t=0}^{t_m}\left\{\widehat{CIF}_{ji}(t;x_1,x_2) - CIF_j(t;x_1,x_2)\right\}^2$$

is approximated by the following sum over 200 equally spaced time points in $[0, t_m]$

$$\frac{t_m}{200}\sum_{t\in\left\{0,\frac{t_m}{200},\ldots,\frac{199}{200}t_m\right\}}\left\{\widehat{CIF}_{ji}(t;x_1,x_2) - CIF_j(t;x_1,x_2)\right\}^2$$

Second, let $H(x_1,x_2)$ denote the expressions inside the integrals $\int_{x_2}$ of (4.5.3), then for each value of x_1 (0 or 1) $\int_{x_2} H(\cdot,x_2)\,dF(x_2)$ is approximated by taking the average of $H(x_1,\cdot)$ evaluated at 150 random values generated from the distribution of X_2. Let $\mathbf{X}_{2,x_1}$ be the set of such values for a specific x_1. Accordingly the final approximation of (4.5.3), the so-called Monte Carlo mean of average integrated square error (MCSE), is

$$\begin{aligned} MCSE_j &= \frac{t_m\,(150\times 200)^{-1}}{2N_S} \\ &\times \sum_{i=1}^{N_S}\left[\sum_{x_1\in\{0,1\}}\sum_{x_2\in\mathbf{X}_{2,x_1}}\left(\sum_{t\in\left\{0,\frac{t_m}{200},\ldots,\frac{199}{200}t_m\right\}}\left\{\widehat{CIF}_{ji}(t;x_1,x_2) - CIF_j(t;x_1,x_2)\right\}^2\right)\right] \end{aligned} \quad (4.5.4)$$

The MAISE and its approximation, the MCSE, measure the accumulated square error of the estimator over time and over covariate distributions. Of note, once generated, the same values of X_2 in $\mathbf{X}_{2,x_1}$ was used for calculating the $MCSE_j$ for all compared methods. This reduces variability in the difference or ratio between the $MCSE_j$ from different methods.

4.5.7 Results

Results for all three mixture scenarios, namely mixture Weibull (2×Weibull), two SNP stdnorm (2×SNPN) and two logmixturenormal (2×logmixturenorm), are tabulated in Tables 4.2, 4.3, 4.4, 4.5 and 4.6. In these tables, the first 4 settings of each scenario correspond to combinations of 10% and 50% right-censoring with $n = 100$ and $n = 500$, and the last 2 settings correspond to interval-censoring with 10% right-censoring. Results from Fine and Gray, and CSH scenarios are reported in Table 4.7.

In the 2×SNPN scenario with $n = 100$, 21% (42/200) of the SNP fits had a non-positive definite Hessian indicating issues with the estimation algorithm. However, this also happened to 13% and 7% of results from the parametric mixture lognormal and mixture Weibull methods, respectively, in the same setting. A closer look at this setting reveals that subjects having the first competing risk had events very early as reflected in the left panel of Figure 4.5.2. Consequently, among all subjects having the first competing risk from the setting, more than 60% had event times within the first interval $[0, t_m/12]$. This and the fact that the sample size of $n = 100$ is relatively low might lead to inadequate information to fit even a simple parametric mixture model to these data. Due to this unexpected phenomenon, results from this whole simulation setting are not reported. Moreover, as mentioned in Subsection 4.5.5 Fine and Gray method and Cox proportional CSH method cannot deal with interval-censoring. Thus no results from these methods for interval-censored data were obtained.

Mixture scenarios with only right-censoring

Among a total of 2400 right-censored data sets simulated according to the mixture scenarios, my estimation algorithm for the SNP models resulted in 12 (0.5%) fits with ill-behaved (i.e. not positive definite) Hessians. Most (10) of these cases were observed in the two last settings: 2×logmixturenorm scenarios with 50% right-censoring (and $n = 100$ with 4 cases or $n = 500$ with 6 cases), while the other two came from a 2×Weibull setting with 50% right-censoring and $n = 500$, and a 2×SNPN setting having 10% right-censoring and $n = 500$. The parametric mixture Weibull method also resulted in one fit with an ill-behaved covariance in a 2×logmixturenorm setting where there is 50% right-censoring and $n = 100$. All data sets involving ill-behaved covariance matrices resulting from any of the compared methods were excluded in all simulation reports.

Monte Carlo (MC) means of SNP estimates are shown in Table 4.2. In most settings average estimates of the regression parameters in the AFT models as well as non-intercept terms in the multinomial logistic model agreed closely with the true values indicating low bias. One exception is the mixture Weibull setting with 50% right-censoring and sample size of 100 where estimates of $\beta_{11}, \beta_{21}, \gamma_0$ and γ_1 are noticeably inaccurate. Finally from Table 4.2 it can also be seen that bias and variability of all

parameter estimates tended to worsen as the amount of right-censoring increased, especially for small sample size.

Tables 4.3 and 4.4 show the relative mean square error (MSE) of parameter estimates from the SNP models compared to parametric models. Values of relative MSE larger than one favour the SNP results. For mixture Weibull scenarios, estimates of regression parameters from the mixture Weibull model lead to smaller MSE than the SNP results though the benefit of fitting the true parametric model was less pronounced for the larger sample size. Performance of the log-normal parametric model in these scenarios lead to larger MSE than the SNP model for most parameters, especially for the larger sample size. For the 2×SNPN scenarios, AFT parameters were estimated more accurately by the SNP method compared to parametric model except for the AFT for event type 1 (which was simulated according to a lognormal distribution), where the lognormal model performed best.

Parameters for the multinomial logistic model tended to be more accurately estimated by both parametric methods compared to the SNP model in the 2×SNPN scenarios for $n = 100$, but for $n = 500$ MSEs for all 3 models were very similar. The benefit of the SNP model was most pronounced for the 2×logmixturenorm scenarios where it dramatically outperformed the Weibull model regardless of sample size and also showed a clear advantage over the lognormal model for $n = 500$. Of note, the big increase in relative efficiency of β_{21} between $n = 100$ and $n = 500$ of the scenario 2×logmixture with 50% right-censoring in both Tables 4.3 and 4.4 can be largely explained by the massive (tenfold) increase in precision (MC-SD) for the SNP estimator, see the corresponding cells in Table 4.2.

Table 4.5 compares the accuracy of covariate-dependent CIF estimates (as measured by the relative MCSE) between my SNP model and alternative parametric and semi-parametric methods. A relative MCSE larger than one supports the SNP method. Overall, my SNP model performed better than parametric alternatives, as it was never substantially worse (the lowest relative MCSE was 0.93) but performed substantially better than parametric methods for some scenarios (the largest relative MCSE were > 3 for both parametric models). Moreover, the SNP model substantially outperformed the semi-parametric Fine and Gray and Cox proportional CSH models across all mixture scenarios. Of note, for the mixture Weibull settings, CIF estimation from the Cox proportional CSH method was extremely unstable resulting in out of range estimates. This might be due to a corrupt estimate of the total survival time based on the Nelson-Aalen estimator for the cumulative cause-specific hazards at timepoints with small risk sets, or because of covariate values that are extreme (see page 117 of Beyersmann et al. (2012)). Consequently, these results were not reported.

The observed coverage of asymptotic 95% confidence intervals of my SNP model is reported in Table 4.6. In all settings, for the lower sample size of $n = 100$, considerable undercoverage was observed, especially for AFT parameter estimates. The lowest observed coverage was 57.5%. Moreover,

only for this sample size, a clear adverse effect on 95%-CI coverage of increasing right-censoring was seen. When sample size was increased to $n = 500$, 95%-CI coverage of all parameter estimates was substantially improved. In many cases the nominal coverage was achieved. However, some undercoverage remained and the lowest observed level was 87%. For mixture Weibull and 2×SNPN scenarios, which correspond to SNP models, one potential explanation for this undercoverage is that the SNP fits based on AIC frequently did not identify the correct base-K i.e. the correct combination of base density and the polynomial degrees K_1 and K_2. This is in line with results shown in Table 3.2 for the situation without covariates. By contrast, model selection based on BIC was much more successful in identifying the correct base-K. However, this did not unfortunately improve coverage compared to the results reported in Table 4.6.

Table 4.2: Accuracy and precision of SNP estimation in mixture scenarios.

Censoring	n	β_{11} = -2	β_{12} = 1	β_{21} = -1	β_{22} = 0	γ_0 = -0.71	γ_1 = -0.5	γ_2 = 1
		MC mean and MC standard deviation						
2×Weibull								
10% RC	100	**-2.03** 0.62	**0.90*** 0.57	**-1.00** 0.06	**0.00** 0.07	**-0.72** 0.33	**-0.49** 0.49	**1.04** 0.53
	500	**-2.01** 0.21	**1.00** 0.20	**-1.00** 0.03	**0.00** 0.02	**-0.69** 0.15	**-0.52** 0.20	**1.00** 0.25
50% RC	100	**-2.31*** 1.03	**0.99** 0.84	**-0.94*** 0.20	**-0.01*** 0.10	**-0.33*** 0.98	**-0.93*** 1.13	**1.10** 0.80
	500	**-2.03** 0.41	**0.97** 0.30	**-0.99*** 0.06	**0.00** 0.04	**-0.66** 0.37	**-0.56*** 0.41	**1.02** 0.30
RC+ IC	100	**-2.52*** 2.61	**0.95** 0.68	**-1.00** 0.08	**0.00** 0.08	**-0.65*** 0.44	**-0.65*** 0.60	**1.05** 0.61
	500	**-2.02** 0.28	**0.97** 0.25	**-1.00** 0.03	**0.00** 0.03	**-0.72** 0.19	**-0.51** 0.25	**1.01** 0.25
2×SNPN								
10% RC	100	**-2.03** 0.45	**1.00** 0.47	**-0.99** 0.18	**-0.03*** 0.17	**-0.73** 0.35	**-0.53** 0.53	**1.11*** 0.59
	500	**-2.00** 0.16	**1.02** 0.16	**-1.00** 0.06	**0.01** 0.06	**-0.72** 0.16	**-0.49** 0.24	**1.03** 0.23
50% RC	100	**-2.05** 0.55	**1.06** 0.60	**-0.96** 0.48	**-0.06*** 0.33	**-0.62** 0.64	**-0.60*** 0.69	**1.13*** 0.66
	500	**-1.99** 0.18	**1.00** 0.20	**-1.00** 0.08	**0.00** 0.08	**-0.70** 0.20	**-0.50** 0.25	**1.03** 0.26
RC+ IC	500	**-2.25*** 1.12	**1.01** 0.29	**-0.99** 0.07	**-0.01** 0.08	**-0.69*** 0.14	**-0.53*** 0.21	**1.05*** 0.22
2×logmixturenorm								
10% RC	100	**-1.99** 0.37	**0.99** 0.54	**-1.00** 0.07	**0.00** 0.07	**-0.73** 0.38	**-0.51** 0.47	**1.09*** 0.58
	500	**-2.04*** 0.16	**0.99** 0.15	**-1.00*** 0.02	**0.00** 0.02	**-0.67*** 0.15	**-0.55*** 0.22	**1.01** 0.24
50% RC	100	**-2.07** 0.52	**0.98** 0.56	**-0.93*** 0.23	**0.01** 0.09	**-0.58*** 0.79	**-0.75*** 0.89	**1.06** 0.82
	500	**-1.99** 0.19	**1.02** 0.19	**-0.99*** 0.02	**0.00** 0.03	**-0.69** 0.29	**-0.52** 0.33	**1.05*** 0.28
RC+ IC	100	**-2.08*** 0.58	**1.07** 0.56	**-0.96*** 0.07	**0.00** 0.06	**-0.68** 0.43	**-0.58*** 0.53	**1.07** 0.56
	500	**-2.11*** 0.29	**1.00** 0.18	**-0.96*** 0.03	**0.00** 0.02	**-0.59*** 0.18	**-0.64*** 0.22	**1.01** 0.23

Note: RC is right censoring, IC is interval censoring.
*: true parameter is not contained in the interval MC mean $\pm$ 1.96 $\times$ (MC SD) $/\sqrt{m}$, $m = 200$.

Table 4.3: Relative efficiency (ratio of MC MSE) between mixture Weibull and SNP results.

		Relative MSE Weibull / SNP and bootstrapped SE						
Censoring	n	β_{11}	β_{12}	β_{21}	β_{22}	γ_0	γ_1	γ_2
2×Weibull								
10% RC	100	**0.66**	**0.94**	**0.92**	**0.86**	**1.02**	**1.02**	**1.00**
		0.06	0.07	0.05	0.06	0.04	0.02	0.01
	500	**0.93**	**0.97**	**0.98**	**0.94**	**1.02**	**1.02**	**1.00**
		0.03	0.03	0.02	0.02	0.02	0.02	0.00
50% RC	100	**0.75**	**0.72**	**0.86**	**0.82**	**0.93**	**0.94**	**0.99**
		0.07	0.06	0.13	0.07	0.10	0.09	0.04
	500	**0.90**	**0.86**	**0.86**	**0.92**	**0.83**	**0.86**	**1.00**
		0.04	0.05	0.07	0.04	0.09	0.07	0.01
RC+IC	100	**0.97**	**0.82**	**0.70**	**0.78**	**0.90**	**0.96**	**1.00**
		0.02	0.09	0.08	0.05	0.05	0.05	0.03
	500	**0.76**	**0.87**	**0.93**	**0.95**	**0.93**	**0.95**	**0.97**
		0.16	0.03	0.03	0.02	0.03	0.03	0.01
2×SNPN								
10% RC	100	**0.92**	**1.02**	**1.44**	**1.46**	**0.92**	**0.96**	**0.97**
		0.12	0.19	0.14	0.15	0.05	0.02	0.02
	500	**1.38**	**1.47**	**2.11**	**1.75**	**1.00**	**0.99**	**0.97**
		0.11	0.14	0.21	0.17	0.01	0.01	0.01
50% RC	100	**1.08**	**0.98**	**0.99**	**1.29**	**0.66**	**0.82**	**0.90**
		0.13	0.12	0.22	0.19	0.11	0.11	0.05
	500	**1.82**	**1.46**	**4.59**	**2.41**	**0.95**	**0.99**	**1.02**
		0.16	0.14	0.59	0.24	0.05	0.03	0.02
RC+IC	500	**0.40**	**0.84**	**2.09**	**1.24**	**0.92**	**0.98**	**0.97**
		0.08	0.11	0.20	0.10	0.04	0.02	0.01
2×logmixturenorm								
10% RC	100	**1.95**	**1.19**	**5.02**	**4.88**	**1.24**	**1.15**	**1.06**
		0.23	0.12	0.84	0.99	0.08	0.08	0.03
	500	**5.21**	**2.29**	**10.46**	**10.26**	**1.45**	**1.10**	**1.10**
		0.63	0.23	1.37	1.53	0.14	0.09	0.03
50% RC	100	**1.60**	**1.38**	**2.84**	**6.68**	**1.70**	**1.56**	**1.10**
		0.17	0.15	0.83	1.27	0.16	0.14	0.05
	500	**4.26**	**2.95**	**197.14**	**13.59**	**5.56**	**4.59**	**1.18**
		0.50	0.36	27.15	2.26	0.70	0.57	0.09
RC+IC	100	**0.93**	**1.18**	**6.74**	**6.06**	**0.96**	**0.95**	**1.08**
		0.13	0.14	1.00	0.93	0.07	0.07	0.04
	500	**0.98**	**1.81**	**2.84**	**7.52**	**0.69**	**0.74**	**1.09**
		0.27	0.22	0.33	0.93	0.06	0.05	0.04

Note: RC is right censoring, IC is interval censoring.

Table 4.4: Relative efficiency (ratio of MC MSE) between mixture lognormal and SNP results.

		Relative MSE Lognormal / SNP and bootstrapped SE						
Censoring	n	β_{11}	β_{12}	β_{21}	β_{22}	γ_0	γ_1	γ_2
2×Weibull								
10% RC	100	**0.84**	**1.06**	**1.26**	**1.16**	**0.96**	**0.99**	**0.99**
		0.06	0.09	0.12	0.10	0.04	0.03	0.02
	500	**1.19**	**1.55**	**1.44**	**1.52**	**0.95**	**0.96**	**1.04**
		0.10	0.14	0.12	0.13	0.05	0.04	0.03
50% RC	100	**1.19**	**0.90**	**1.11**	**1.10**	**1.24**	**1.14**	**1.03**
		0.12	0.10	0.11	0.15	0.11	0.08	0.06
	500	**2.04**	**1.43**	**1.74**	**1.28**	**2.14**	**1.81**	**1.05**
		0.26	0.16	0.35	0.14	0.43	0.30	0.05
RC+IC	100	**0.90**	**0.84**	**0.87**	**0.96**	**0.84**	**0.89**	**1.01**
		0.02	0.12	0.10	0.08	0.04	0.04	0.04
	500	**0.77**	**0.95**	**1.32**	**1.25**	**0.84**	**0.91**	**0.93**
		0.16	0.06	0.11	0.09	0.04	0.04	0.02
2×SNPN								
10% RC	100	**0.74**	**0.64**	**2.16**	**2.11**	**1.00**	**0.98**	**0.96**
		0.06	0.10	0.25	0.24	0.04	0.02	0.02
	500	**0.97**	**0.91**	**3.63**	**2.84**	**1.01**	**1.00**	**0.99**
		0.03	0.04	0.40	0.27	0.01	0.01	0.01
50% RC	100	**0.71**	**0.64**	**1.40**	**1.90**	**0.70**	**0.81**	**0.93**
		0.07	0.05	0.25	0.30	0.10	0.08	0.05
	500	**0.96**	**0.98**	**8.65**	**4.47**	**0.96**	**1.01**	**1.03**
		0.02	0.03	1.15	0.51	0.03	0.01	0.01
RC+IC	500	**0.83**	**0.73**	**1.88**	**1.53**	**0.93**	**0.99**	**0.99**
		0.09	0.09	0.20	0.14	0.04	0.02	0.01
2×logmixturenorm								
10% RC	100	**0.87**	**0.59**	**4.39**	**4.38**	**1.19**	**1.13**	**1.03**
		0.09	0.07	0.72	0.85	0.07	0.06	0.03
	500	**1.21**	**1.11**	**9.62**	**11.44**	**1.44**	**1.08**	**1.07**
		0.14	0.07	1.41	1.58	0.15	0.09	0.03
50% RC	100	**0.83**	**0.77**	**0.67**	**4.22**	**1.05**	**1.00**	**1.12**
		0.08	0.09	0.23	0.86	0.08	0.08	0.06
	500	**1.26**	**1.04**	**22.65**	**9.63**	**1.43**	**1.35**	**1.09**
		0.10	0.06	3.06	1.58	0.13	0.12	0.05
RC+IC	100	**0.68**	**0.87**	**5.32**	**5.82**	**0.88**	**0.91**	**1.08**
		0.09	0.09	0.71	0.91	0.06	0.06	0.04
	500	**0.43**	**0.90**	**2.12**	**7.69**	**0.62**	**0.72**	**1.06**
		0.10	0.07	0.25	0.99	0.04	0.04	0.03

Note: RC is right censoring, IC is interval censoring.

Table 4.5: Accuracy of covariate-dependent CIF estimates, as measured by the relative MCSE (ratio of MCSE) of the best SNP fits compared to alternative models in mixture scenarios.

Censoring	n	**Relative MCSE** and bootstrapped SE							
		Weibull/SNP		Lognormal/SNP		FG/SNP		CSH/SNP	
		CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2
2×Weibull									
10% RC	100	**0.95**	**0.97**	**0.96**	**1.09**	**1.25**	**5.25**	-	-
		0.01	0.01	0.01	0.02	0.03	0.25	-	-
	500	**0.99**	**0.99**	**1.09**	**1.50**	**2.98**	**22.10**	-	-
		0.00	0.00	0.02	0.04	0.14	1.40	-	-
50% RC	100	**0.97**	**0.93**	**0.94**	**1.06**	**1.04**	**1.54**	-	-
		0.01	0.01	0.02	0.03	0.02	0.07	-	-
	500	**0.98**	**0.98**	**1.07**	**1.61**	**2.22**	**4.13**	-	-
		0.01	0.01	0.02	0.06	0.07	0.19	-	-
RC+IC	100	**0.95**	**0.97**	**0.96**	**1.09**	-	-	-	-
		0.01	0.01	0.01	0.02	-	-	-	-
	500	**0.99**	**0.99**	**1.09**	**1.50**	-	-	-	-
		0.00	0.00	0.02	0.04	-	-	-	-
2×SNPN									
10% RC	100	**1.00**	**1.09**	**0.98**	**1.26**	**1.10**	**1.51**	**1.08**	**1.18**
		0.01	0.02	0.01	0.04	0.03	0.05	0.02	0.03
	500	**1.03**	**1.44**	**1.00**	**2.67**	**2.00**	**4.41**	**1.85**	**1.58**
		0.00	0.04	0.00	0.11	0.07	0.19	0.06	0.05
50% RC	100	**0.99**	**1.07**	**0.94**	**1.41**	**1.15**	**1.22**	**1.12**	**1.23**
		0.01	0.04	0.02	0.07	0.04	0.05	0.04	0.05
	500	**1.06**	**1.75**	**1.00**	**4.76**	**2.66**	**2.01**	**2.41**	**1.93**
		0.01	0.06	0.00	0.23	0.12	0.08	0.10	0.07
RC+IC	500	**0.98**	**0.98**	**1.07**	**1.61**	-	-	-	-
		0.01	0.01	0.02	0.06	-	-	-	-
2×logmixturenorm									
10% RC	100	**1.04**	**1.52**	**0.97**	**1.53**	**1.42**	**2.00**	**1.15**	**1.22**
		0.02	0.03	0.02	0.04	0.04	0.07	0.03	0.04
	500	**1.30**	**3.15**	**1.05**	**3.27**	**4.00**	**5.59**	**2.33**	**1.99**
		0.03	0.10	0.02	0.11	0.18	0.23	0.09	0.07
50% RC	100	**1.05**	**1.66**	**0.99**	**1.51**	**1.26**	**1.41**	**1.18**	**1.28**
		0.02	0.06	0.02	0.06	0.04	0.05	0.03	0.05
	500	**1.28**	**4.15**	**1.02**	**3.79**	**3.32**	**3.25**	**2.60**	**2.32**
		0.03	0.12	0.02	0.10	0.12	0.10	0.09	0.07
RC+IC	100	**1.00**	**1.09**	**0.98**	**1.26**	-	-	-	-
		0.01	0.02	0.01	0.04	-	-	-	-
	500	**1.03**	**1.44**	**1.00**	**2.67**	-	-	-	-
		0.00	0.04	0.00	0.11	-	-	-	-

Note: RC is right censoring, IC is interval censoring. -: not available.

Table 4.6: Observed coverage of asymptotic 95% confidence intervals for the SNP estimates.

Censoring	n	SNP method						
		β_{11}	β_{12}	β_{21}	β_{22}	γ_0	γ_1	γ_2
2×Weibull								
10% RC	100	71.0	70.5	94.0	86.5	95.5	98.0	97.0
	500	94.5	93.5	94.0	96.0	95.5	98.0	94.0
50% RC	100	57.5	68.0	72.0	85.0	80.0	81.5	93.5
	500	87.0	89.0	96.0	94.5	92.0	92.0	93.0
RC+IC	100	63.0	62.5	78.0	75.5	85.5	92.0	93.0
	500	93.5	91.5	93.5	94.0	92.0	94.5	93.5
2×SNPN								
10% RC	100	70.0	69.0	88.5	88.0	92.5	93.0	92.5
	500	92.5	95.0	92.0	94.0	92.5	91.0	93.5
50% RC	100	63.5	59.5	81.0	79.5	91.5	95.5	94.0
	500	93.0	93.5	95.0	96.0	94.0	93.5	96.5
RC+IC	500	89.0	88.5	94.5	93.5	96.5	93.0	93.5
2×logmixturenorm								
10% RC	100	75.5	69.0	86.5	89.0	98.0	96.5	95.0
	500	90.5	92.0	88.0	92.5	95.0	92.5	93.0
50% RC	100	72.0	69.5	89.0	89.5	86.0	86.5	91.5
	500	90.0	88.5	96.0	87.0	93.0	92.5	94.0
RC+IC	100	71.0	71.0	83.0	91.5	89.0	93.5	93.0
	500	92.0	91.5	62.0	93.0	86.5	90.5	94.0

Note: RC is right censoring, IC is interval censoring.
Estimated standard error of Monte Carlo coverage entries $\approx 1.5\%$.

Mixture scenarios with interval- and right-censoring

From Table 4.2, compared to settings with only 10% right-censored data, the introduction of interval censoring to these settings generally increased bias and variability in the estimates of all regression parameters. Marked differences were observed in settings with $n = 100$.

Table 4.3 shows that in mixture Weibull settings MSEs of all parameter estimates by the mixture Weibull method are significantly smaller than those from the SNP method. In these settings, most SNP estimates also had larger MSE than those from the mixture lognormal method, especially the multinomial logistic parameters. This was also the case for the 2×SNPN scenarios with $n = 500$, where both parametric mixture methods outperformed the SNP method in terms of MSE except for the estimates of β_{21} and β_{22}. In 2×logmixturenorm scenarios, SNP estimates for the AFT parameters were comparable to or much better than those from the mixture Weibull method. However this did not hold for the AFT parameters of the first competing risk when comparing SNP method to mixture lognormal method. SNP estimates of multinomial logistic parameters had larger MSE than those from both mixture Weibull and mixture lognormal methods.

In terms of MCSE (Table 4.5), the mixture Weibull method was slightly better than the SNP method in mixture Weibull settings, whereas the advantage of SNP method over the mixture lognormal method was only noticeable for the setting with large sample size. For the 2×SNPN setting with $n = 500$, SNP method were comparable to or more favorable than both parametric mixture methods. The most pronounced benefit of the SNP method was observed in the 2×logmixturenorm setting with $n = 500$.

Similar to the simulation of mixture scenarios with only right-censoring, noticeable undercoverage of SNP estimates was also observed for interval-censored data with the lower small sample size of $n = 100$. Increasing sample size improved the coverage, in some cases up to the nominal level, but some marked undercoverage could still be seen.

Fine and Gray scenarios and CSH scenarios

From Table 4.7, there was in general no clear benefit of using SNP method over the parametric mixture methods with respect to accuracy of covariate–dependent CIF estimates. This is not unexpected as Figures 4.5.5 and 4.5.4 indicate that the true CIFs from the Fine and Gray, and Cox CSH scenarios appear to be well approximated by simple parametric curves.

For the Fine and Gray scenarios, the Fine and Gray method was slightly better than the SNP method in estimating the first CIF, which exactly followed a Fine and Gray model. However it lost to SNP method in estimating the other CIF. In these settings, MCSEs from SNP CIF estimates were often smaller than those from the Cox CSH method. Finally in CSH scenarios, the Cox CSH method tended to surpass the SNP method. The SNP method was slightly better than the Fine and Gray method in estimating the first CIF, whereas the opposite was true for the estimation of the second CIF.

Table 4.7: Accuracy of covariate-dependent CIF estimates, as measured by the relative MCSE (ratio of MCSE), of the SNP model compared to alternative models in Fine and Gray and CSH scenarios.

		Relative MCSE Bootstrapped SE							
Censoring	n	Weibull/SNP		Lognormal/SNP		FG/SNP		CSH/SNP	
		CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2	CIF_1	CIF_2
Fine and Gray									
10% RC	100	**0.97**	**1.00**	**0.96**	**0.95**	**0.94**	**1.31**	**1.08**	**1.16**
		0.01	0.00	0.01	0.00	0.02	0.01	0.04	0.01
	500	**1.00**	**1.00**	**1.08**	**0.95**	**0.94**	**1.37**	**1.97**	**1.18**
		0.00	0.00	0.02	0.00	0.01	0.01	0.08	0.00
50%RC	100	**0.95**	**1.00**	**0.86**	**0.89**	**0.88**	**1.16**	**0.93**	**1.11**
		0.02	0.01	0.02	0.01	0.03	0.02	0.03	0.01
	500	**0.94**	**1.00**	**1.04**	**0.89**	**0.93**	**1.26**	**1.31**	**1.15**
		0.01	0.00	0.03	0.00	0.02	0.01	0.04	0.01
CSH									
10%RC	100	**0.98**	**1.00**	**1.03**	**0.94**	**0.98**	**1.10**	**0.98**	**1.01**
		0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01
	500	**1.00**	**1.00**	**1.18**	**0.94**	**1.05**	**1.11**	**1.04**	**1.02**
		0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
50%RC	100	**0.93**	**0.99**	**0.97**	**0.86**	**0.79**	**1.03**	**0.84**	**1.00**
		0.01	0.01	0.02	0.01	0.02	0.01	0.02	0.01
	500	**0.99**	**1.00**	**1.31**	**0.84**	**0.86**	**1.09**	**0.95**	**1.02**
		0.01	0.00	0.02	0.00	0.01	0.01	0.01	0.01

4.5.8 Summary

Simulation results for regression are more difficult to interpret and less in favour of the SNP method compared to the results for CIF estimation reported in Chapter 3. Nevertheless, some advantages of the SNP approach compared to alternative methods were observed: First, for mixture scenarios, parameter estimates can be substantially more precise than results from parametric mixture models if the truth does not follow a simple parametric model. Second, covariate-specific CIF predictions based on SNP models were overall competitive compared to all investigated alternative parametric and semi-parametric models. Indeed, compared to each alternative model, performance of the SNP model was substantially better for at least one simulation scenario and never dramatically worse.

Asymptotic confidence intervals showed clear undercoverage for many parameter estimates across all mixture settings for a small sample size of $n = 100$. This indicates unreliability of standard MLE-based inference in cases with sample size (and censoring proportions) of similar magnitude. Nevertheless, the simulation also shows that with a five-fold increase in sample size one could potentially rely on standard MLE-based inference.

4.6 Model checking in competing risks

From the previous sections, the final fit from a SNP model is a parametric model. Accordingly the validity of this final model can be informally verified by means of diagnostic approaches for parametric competing risks models. Despite a wealth of research on model diagnostics for standard survival analysis, see Collett (2003) for a summary, to my knowledge there has not been any systematic discussion on model diagnostics for competing risks models. Thus, in this section a short exposition of model diagnostics for competing risks models will be given.

4.6.1 Model assumptions in competing risks

To begin with, the basic assumptions for parametric competing risks model with time-independent covariates are:

1. The competing risks process underlying the observed data is adequately described by the parametric model.
2. The observations are independent of each other.
3. Conditional on covariates $\mathbf{X}$, the censoring (truncation) mechanism is independent of the competing risks process.

Assumption 2 is often not examined as it is implied by the context of data collection but can be informally checked by plotting residuals against a (measurable) factor that is suspected of inducing some dependency between the observations e.g. inclusion times of the subjects. In well-conducted studies with a pre-defined follow-up schedule and little loss to follow-up, assumption 3 is automatically fulfilled. Assumption 1 can be formally and informally checked via statistical tests and graphical tools and is the main topic of this section. In what follows I shall mainly focus on right-censored data which will be followed by a brief discussion on extension to interval-censored data, whereas the case of left-truncation will not be covered.

4.6.2 Model checking for CIF estimation

If there are no or only a few categorical covariates, the appropriateness of a parametric model for the CIFs can be visually assessed by comparing the (stratum-specific) parametric CIF estimates to their nonparametric counterparts and corresponding point-wise 95%-confidence intervals. This can be supplemented with a test of the null hypothesis that the nonparametric CIF estimate is compatible with the fitted parametric model. If the test shows no statistical evidence for rejecting the null hypothesis and the graph also shows no discrepancy then one may confidently accept the chosen model.

One option for such a test is to apply the method described in Hollander & Proschan (1979) for right-censored survival data to the situation of CIF estimation. Specifically, assume that there are two competing risks and that the relevant null hypothesis that the true CIF describing the data at hand (CIF_1) has a known parametric form $CIF_{1\theta}$ i.e. $H_0 : CIF_1(\cdot) \equiv CIF_{1\theta}(\cdot)$, let X and Y be the improper random variables with cumulative distributions given by CIF_1 and $CIF_{1\theta}$ and point masses at time infinity of size $1 - CIF_1(\infty)$ and $1 - CIF_{1\theta}(\infty)$, respectively. Under H_0, $\int_0^{t_{\max}} CIF_1(y)dCIF_{1\theta}(y) = \frac{1}{2}CIF_{1\theta}^2(t_{\max})$, where $t_{\max}$ is the maximum observed event time. Thus, under H_0

$$CIF_{1\theta}^{-2}(t_{\max}) \int_0^{t_{\max}} CIF_1(y)dCIF_{1\theta}(y) = \frac{1}{2}$$

which can also be interpreted as $P(X < Y | X < t_{\max}, Y < t_{\max})$. Hence the appropriate test statistic is

$$C_{t_{\max}} = CIF_{1\theta}^{-2}(t_{\max}) \int_0^{t_{\max}} \widehat{CIF}_1(y)d\widehat{CIF}_{1\theta}(y) \tag{4.6.1}$$

where $\widehat{CIF}_1$ is the nonparametric estimator for CIF_1 and $CIF_{1\theta}$ is the parametric CIF estimate from the model being diagnosed. If the uncertainty in estimating $CIF_{1\theta}$ is ignored, i.e. it is assumed to be a priori known and fixed, then $C_{t_{\max}}$ has an asymptotic normal distribution $\mathcal{N}(1/2, \sigma)$ under H_0 and a consistent variance estimator is available as shown in Appendix B.1. To avoid the complex estimation of the variance of $C_{t_{\max}}$, a p-value can also be derived by following the resampling technique of Lin (1997).

As acknowledged in Hollander & Proschan (1979), one limitation of this test is that it is not sensitive to alternatives leading to $P(X < Y | X < t_{\max}, Y < t_{\max}) = \frac{1}{2}$. Moreover, the consequences of ignoring the fact that the parametric model itself is also estimated from the observed data in the derivation of the asymptotic distribution are somewhat unclear, though the test might still have value as an informal tool supplementing the graphical analysis. A possible work-around for this latter problem is to bootstrap the test statistic by repeatedly fitting both the parametric and nonparametric models to bootstrap samples to obtain a more accurate estimate of the standard error of the test statistic. As this approach would be computationally intensive and its validity is difficult to establish formally, I have not pursued it further.

4.6.3 Diagnostics for models based on mixture factorization

When there is no censoring, one can use separate diagnostic procedures for each of the $J + 1$ component models in a competing risks model based on the mixture factorization: The J AFT models for $P(T \mid D = j)$ and the multinomial logistic model for $P(D)$. This offers the use of standard methods available for each submodel. Note that the problem at hand is not right-censoring per se, as standard

model diagnostics for AFT models can cope with right-censoring, but the fact that the event type D is also unknown for right-censored observations.

To circumvent this problem, I suggest the following imputation-based approach: For each right-censored observation (with censoring time t_{cens}), impute the event type D by drawing it from the multinomial distribution with cell probabilities $P(D = j \mid T > t_{cens}; \mathbf{X}), j = 1, \ldots, J$, which can be derived from the fitted model. Then all observed and imputed event types are used for diagnosing the multinomial logistic model. In addition, the AFT model for each event type j is diagnosed by using observed event times of type j and right-censored times from observations with imputed event type j. There are, however, some issues with the imputation:

1. It makes the model potentially look too good as one simulates from the model, which one wants to check.

2. Diagnostic plots might be affected by the specific random draw, i.e. they might look different if the random imputation is repeated.

These issues can be informally addressed by plotting imputed observations in a different colour in all diagnostic plots and by repeating these plots for multiple imputed data sets. An additional issue is that when censoring is heavy, the diagnostic plots are dominated by the imputed points. However, as discussed Section 4.4, heavy right-censoring is often associated with a short follow-up duration relative to event times which causes identifiability problems for models based on the mixture-factorization model. Thus, in this case one should consider abandoning any model based on the mixture-factorization altogether.

Model diagnostics for multinomial logistic models for $\mathbf{P(D = j)}$

When $P(D)$ is modelled by a multinomial logistic model, diagnostic tools designated for this type of model as discussed in Agresti (2002) can be employed. In the current setting, for each competing event j, the multinomial setting is simplified to a binomial setting by considering only $P(D = j, \gamma; \mathbf{X})$ and $P(D \neq j, \gamma; \mathbf{X})$, where γ and $\mathbf{X}$ denote, respectively, the regression parameters and covariates. When there are only a few categorical covariates, subjects having the same covariate levels can be grouped into one unit yielding data of the form $\{y_{ij}, n_i, \mathbf{X}_i\}_{i=1,\ldots,m}$, $m \leq n$, where n is the sample size, m is the number of groups, and y_{ij} is the number of subjects having event type j amongst the n_i subjects with the same covariate values $\mathbf{X}_i$. Then one can consider the following Pearson-type residual

$$e_{ij} = \frac{y_{ij} - n_i P(D = j, \hat{\gamma}; \mathbf{X}_i)}{\sqrt{n_i P(D = j, \hat{\gamma}; \mathbf{X}_i)\left[1 - P(D = j, \hat{\gamma}; \mathbf{X}_i)\right]}} \tag{4.6.2}$$

For large group size n_i, e_{ij} has an approximate standard normal distribution when $\hat{\gamma}$ is estimated from a correct model. Accordingly a plot of these residuals against the respective linear predictors should show a random scatter around zero without trends for extreme outliers.

For the situation of more than 2 event types ($J > 2$) interpretation of Pearson residuals is more difficult, as the proposed approach generates multiple sets of Pearson residuals all of which depend on the full set of regression coefficients of the multinomial logistic model. Moreover, as the proposed approach involves imputing the event type for right-censored data, the proportion of imputed data as well as the group size n_i should be visually coded in the plot

One complication of interpreting plots based on e_{ij} (even in binomial logistic regression) is that when there are too many strata or when $\mathbf{X}_i$ has continuous elements (causing $n_i = 1$) the residuals plots are difficult to interpret.

Model diagnostics for the AFT models for $\mathbf{P(T \mid D = j)}$

For each event type j, the parametric model for $P(T \mid D = j)$ can be diagnosed by using all subjects observed to have that event type as well as censored observations with imputed event type j, who are treated as right-censored. These can be regarded as an approximate random right-censored sample from the conditional distribution of $T \mid D = j$, a proper survival time with a proper survival function $S_j(t) = P(T > t \mid D = j)$. In the remainder of this section, it is assumed that the event type of interest is j. Let n_j be the number of subjects with observed event type j plus censored subjects whose imputed event type is j; in other words those who were destined (possibly by imputation) to experience event type j. The observed event- or censoring times for those subjects are denoted by $t_{ji}, i = 1, \ldots, n_j$.

A basis for assessing the overall fit of a model for $P(T \mid D = j)$ is the Cox-Snell residual. For an observation at time $t_{ji}, i = 1, \ldots, n_j$, according to Section 7.1.2 of Collett (2003), Cox-Snell residuals are defined as

$$r_{C_{ji}} = -\log \hat{S}_{ji}(t_{ji}) \tag{4.6.3}$$

Subscript i of the estimated survival function indicates its dependence on subject's covariates i.e $\hat{S}_{ji}(t) = \hat{S}_{ji}(t; \mathbf{X}_{ji})$. It is easy to see that regardless of the distribution of T, as long as T is a valid random survival time then $-\log S_T(T)$ always follows a unit exponential distribution (see e.g. Section 4.1.1 of Collett (2003) for a proof). Accordingly $r_{C_{ji}}, i = 1, \ldots, n_j$ becomes a random right-censored sample from a random variable U with a unit exponential distribution if $\hat{S}_{ji}$ is a valid estimate for the survival function of the conditional time-to-event. Consequently, $-\log S_U\left(r_{C_{ji}}\right) = r_{C_{ji}}$ suggesting that if the model is correct, a plot of $-\log \hat{S}_U^{KM}\left(r_{C_{ji}}\right)$ vs. $r_{C_{ji}}$, where $\hat{S}_U^{KM}(\cdot)$ is the Kaplan-Meier

estimator for the survival function of U based on $r_{C_{ji}}$, should roughly resemble a straight line through the origin with unity slope. Any major and systematic deviation from the straight line might question the overall goodness of fit of the relevant AFT model.

When an AFT model: $\log(T \mid D = j) = \mu_j + \mathbf{X}_j^T \beta_j + \sigma_j \epsilon_j(\theta)$, where $\epsilon_j(\theta)$ means the distribution of ϵ_j is parametrized by θ, e.g. a SNP base density, is used

$$\hat{S}_{ji}(t_{ji}) = P\left(\epsilon_j(\theta) > \frac{\log t_{ji} - \hat{\mu}_j - \mathbf{X}_{ji}^T \hat{\beta}_j}{\hat{\sigma}_j}\right) = S_{\epsilon_j(\theta)}\left(r_{S_{ji}}\right)$$

where, as mentioned in Section 7.1.1 of Collett (2003)

$$r_{S_{ji}} = \frac{\log t_{ji} - \hat{\mu}_j - \mathbf{X}_{ji}^T \hat{\beta}_j}{\hat{\sigma}_j} \tag{4.6.4}$$

is the standardized residual. Thus $r_{C_{ji}} = -\log S_{\epsilon_j(\theta)}\left(r_{S_{ji}}\right)$, where $r_{S_{ji}}$ has the form of "expected - observed outcome".

To detect if under the fitted model any event times are unexpectedly large or small, plots of "martingale" residuals vs. observed times or their ranks can be used. Chapter 7 in Andersen et al. (1997) gave theoretical discussions of martingale residuals in competing risks. By definition, martingale residuals are

$$r_{M_{ji}} = \delta_{ji} - r_{C_{ji}} \tag{4.6.5}$$

where $\delta_{ji} = 1$ indicates an event for the i^{th} subject of those bound to have event type j and $\delta_{ji} = 0$ means right-censoring. The corresponding martingale M_{ji} is the difference between the counting process $N_{ji}(t)$, which counts the number of event over time for a subject deemed to have event type j, and the intensity process, which is the product of the conditional hazard $\lambda_{T|D=j}(t, \theta, \mathbf{X}_{ji})$ and the at risk process $Y_i(t)$. Specifically

$$M_{ji} = N_{ji} - \int_0^t Y_i(u) \lambda_{T|D=j}(u, \theta, \mathbf{X}_{ji})\, du \tag{4.6.6}$$

Thus, $r_{M_{ji}}$ has the form of the difference between the observed and expected number of event for a subject in the interval $(0, t_{ji}]$. Moreover, it can be shown that asymptotically $r_{M_{ji}}$ and $r_{M_{jl}}$, $l \neq i$, are uncorrelated, and $E\left(r_{M_{ji}}\right) = 0$. However, this residual is theoretically not symmetric as its range is $(-\infty, 1]$. Thus, a symmetrized modification of the martingale residual called the deviance is used. This residual is defined as

$$r_{D_{ji}} = sgn\left(r_{M_{ji}}\right)\left[-2\left\{r_{M_{ji}} + \delta_{ji} \log\left(\delta_{ji} - r_{M_{ji}}\right)\right\}\right]^{1/2} \tag{4.6.7}$$

which is symmetrically distributed about zero. A (kernel) smoother can be used when plotting martingale or deviance residuals against observed times to trace out systematic deviations or patterns.

Detecting non-linearity and interactions

Non-linearity can be informally assessed by plotting Pearson residuals (for the model for $P(D)$) or standardized residual $r_{S_{ji}}$ (for the AFT models) against covariates given that right-censoring is ignorable.

More formally, to gain insights into the functional form of a covariate for any regression model (including AFT or multinomial logistic models), one can investigate the estimated functional form obtained by including the respective covariate as a flexible function (such as a natural cubic spline with fixed degrees of freedom) rather than a simple linear term into the linear predictor.

To detect interactions of certain covariates in designated components of a model, following standard likelihood setting, one can fit the model with these interactions in the chosen components and use Wald-type tests to verify the strength of the interactions. Alternatively, a likelihood-ratio test comparing the models with and without the interactions can also be used.

4.6.4 Diagnostics for competing risks models in general case

As competing risks modelling is a special case of multi-state modelling, it is expected that some diagnostic tools for multi-state models might be applicable to competing risks models. Titman & Sharples (2010) proposed a summary residual based on comparing the observed and expected states at the observed time. However, this residual is driven by the labelling of the states which are the competing events in competing risks setting. Thus, such an approach could only be appropriate if there exists a natural ordering among the event types which is rare in practice.

Alternatively one can use a martingale-type residual. For right-censored competing risks data, the martingale residual for the ith subject and event type j is

$$r_{\tilde{M}_{ij}} = \mathbf{I}(\delta_i = j) - \int_0^{t_i} \lambda_j \left(u, \hat{\theta}, \mathbf{X}_i\right) du \tag{4.6.8}$$

where $\Lambda_j(t_i) = \int_0^{t_i} \lambda_j \left(u, \hat{\theta}, \mathbf{X}_i\right) du$ is the estimated cumulative cause-specific hazard for event type j. $r_{\tilde{M}_{ij}}$ is the difference between the observed and expected number of event type j over $(0, t_i]$. To avoid confusion with the martingale residual $r_{M_{ji}}$ for the conditional time-to-event sub-models introduced earlier, here the subscript for individual (i) is put before the subscript for competing event (j) and a tilde is used in the notation $\tilde{M}_{ij}$. This means for any event type j, $r_{\tilde{M}_{ij}}$ is computed for all subjects.

$r_{\tilde{M}_{ij}}$ comes from the martingale

$$\tilde{M}_{ij}(t) = N_{ij}(t) - \int_0^{t_i} Y_i(u)\lambda_j(u, \theta, \mathbf{X}_i)\, du \tag{4.6.9}$$

where $N_{ij}(t)$ counts the number of event type j over $(0, t]$ for the i^{th} subject with the corresponding intensity process $Y_i(t)\lambda_j(t, \theta, \mathbf{X}_i)\, dt$. When the fitted model is correct $\tilde{M}_{ij}$ and $\tilde{M}_{ij}$ for $i, j \neq l, k$ are orthogonal martingales Gray (1988). Thus asymptotically, $r_{\tilde{M}_{ij}}$ and $r_{\tilde{M}_{ij}}$ are uncorrelated. As before, due to martingale property $E\left(r_{\tilde{M}_{ij}}\right)$ is asymptotically zero. One way to use these residuals is, for each competing risk j, plot $r_{\tilde{M}_{ij}}$ against the observed times or the respective ranks. When the fitted model is correct, smoothers for all plots should be close to the zero line.

As mentioned before, martingale residuals are in general not symmetrically distributed. Thus, "deviance residuals" should be used, and are defined as

$$r_{\tilde{D}_{ij}} = sgn\left(r_{\tilde{M}_{ij}}\right)\left[-2\left\{r_{\tilde{M}_{ij}} + \mathbf{I}(\delta_i = j)\log\left(\mathbf{I}(\delta_i = j) - r_{\tilde{M}_{ij}}\right)\right\}\right]^{1/2} \tag{4.6.10}$$

4.6.5 Possible extension to interval-censoring

As mentioned in Section 3.6.2, nonparametric CIF estimates no longer give a unique nonparametric estimate for the CIF in the presence of interval-censoring but upper and lower bounds. Nevertheless, in the case of no covariates or a low number of categorical covariates, it is still possible to visually compare the parametric fits with the corresponding estimated CIF bounds generated by the method of Maathuis (2003).

For proportional hazards models for survival data, Farrington (2000) proposed an interesting extensions of residuals for right-censored survival data to interval-censoring and recommended the usage of martingle-type residuals in this context. However, to my knowledge, this approach has not been generalized to AFT survival models with interval-censoring where the properties and applicability of its martingale-type residuals are less clear. Hence I also did not pursue this approach further in the present competing risks setting.

4.6.6 Influential diagnostic for competing risks models

General methods for identifying influential observations in models estimated by MLE can also be applied in the competing risks setting. For reference, the key results are briefly summarized here and are based on Section 3.1 of Titman & Sharples (2010). For a broader discussion which also elaborates on assessing the influence of a subset of observations on a subset of parameters, I refer to Escobar & Meeker (1992).

Individual influence on the whole set of parameters

Let θ be the parameter vector in the chosen parametric model and $\hat{\theta}$, $\hat{\theta}_{(i)}$ be the MLEs of θ for the full data set with n observations and the data set without the ith observation, respectively. To assess the impact of removing the ith observation on the estimate of θ, a natural measure is the Mahalanobis distance between $\hat{\theta}_{(i)}$ and $\hat{\theta}$ i.e.

$$\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^T I\left(\hat{\theta}\right)\left(\hat{\theta}_{(i)} - \hat{\theta}\right) \tag{4.6.11}$$

where $I\left(\hat{\theta}\right)$ is the observed Fisher information evaluated at the MLE. Of note, this quantity can also be motivated as a first-order Taylor approximation to $2\left\{L\left(\hat{\theta}\right) - L\left(\hat{\theta}_{(i)}\right)\right\}$ where L is the log-likelihood function of the full data set. To avoid the computational intensity involving the calculation of this quantity, one can use the following approximation

$$\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^T I\left(\hat{\theta}\right)\left(\hat{\theta}_{(i)} - \hat{\theta}\right) \approx U_i\left(\hat{\theta}, \mathbf{X}_i\right)^T I^{-1}\left(\hat{\theta}\right) U_i\left(\hat{\theta}, \mathbf{X}_i\right) \tag{4.6.12}$$

where $U_i\left(\hat{\theta}, \mathbf{X}_i\right)$ is the score vector or the gradient evaluated at the MLE for the ith observation. From Section 3 of Cain & Lange (1984) this is due to

$$\left(\hat{\theta}_{(i)} - \hat{\theta}\right) \approx I^{-1}\left(\hat{\theta}\right) U_i\left(\hat{\theta}, \mathbf{X}_i\right)$$

This measure of influence can be plotted against the observation number or other quantities of interest, e.g. the observed event or censoring time.

Individual influence on a specific parameter

Let $\hat{\theta}_j$ and $\hat{\theta}_{j(i)}$ be the estimates of the jth parameter in the model from the full data set and the data set without the ith subject, respectively. The obvious measure of influence of this subject on the estimate of θ_j is $\hat{\theta}_{j(i)} - \hat{\theta}_j$ which, according to the previous section, can be approximated by the jth element of the vector $I^{-1}\left(\hat{\theta}\right) U_i\left(\hat{\theta}, \mathbf{X}_i\right)$.

4.7 Applications

In this section, the regression model based on SNP densities formulated in Section 4.1 is applied to the tuberculous meningitis and cryptococcal meningitis data sets introduced in Chapter 3. As both data sets are from randomized clinical trials, I first studied the effect of the intervention as the sole covariate on both the marginal event probabilities and the conditional time-to-event distributions of all competing risks. In a second step, additional baseline covariates were added to the models.

Uncertainty was quantified by 95% confidence intervals which were calculated by two alternative methods: a) using approximate "ad hoc" asymptotic inference as described in Section 3.4, b) using basic bootstrap confidence intervals where, as in Zhang & Davidian (2008), the base density and polynomial degrees where also re-estimated for each bootstrap sample.

As the data set used in the tuberculous example has only right-censoring, diagnostic methods for parametric competing risks model introduced in Section 4.6 are also used to assess the fitted SNP model, once regarded as parametric model.

4.7.1 Initiation of antiretroviral therapy (ART) in HIV–associated tuberculous meningitis (TBM)

As described in Subsection 3.7.1, the outcome of interest for this data set is the time, measured in days, to first neurological event (first competing risk) or prior death (second competing risk). Treatment effect in the simple regression model is represented by a binary covariate with two levels: immediate ART and delayed ART ("placebo"). Parameter estimates from a SNP CIF-based regression model which included treatment as a covariate for all sub-models are shown in Table 4.8. Confidence intervals based on the Hessian matrix of the log-likelihood function (i.e. by "ad hoc" asymptotic inference) were narrower than bootstrap based confidence intervals. There was no evidence that treatment affects any component of the competing risks process. This is in accordance with the result of the IWD-based randomization tests for equality of the CIFs reported in Table 3.14 which also lead to large p-values. Finally, an alternative analysis which fitted Fine and Gray models to each competing event separately also showed that treatment did not significantly affect the absolute risk of either competing events (p-values are 0.94 for neurological event and 0.57 for prior death)

Table 4.8: Estimates of regression coefficients and and corresponding confidence intervals for the model with treatment as the only covariate.

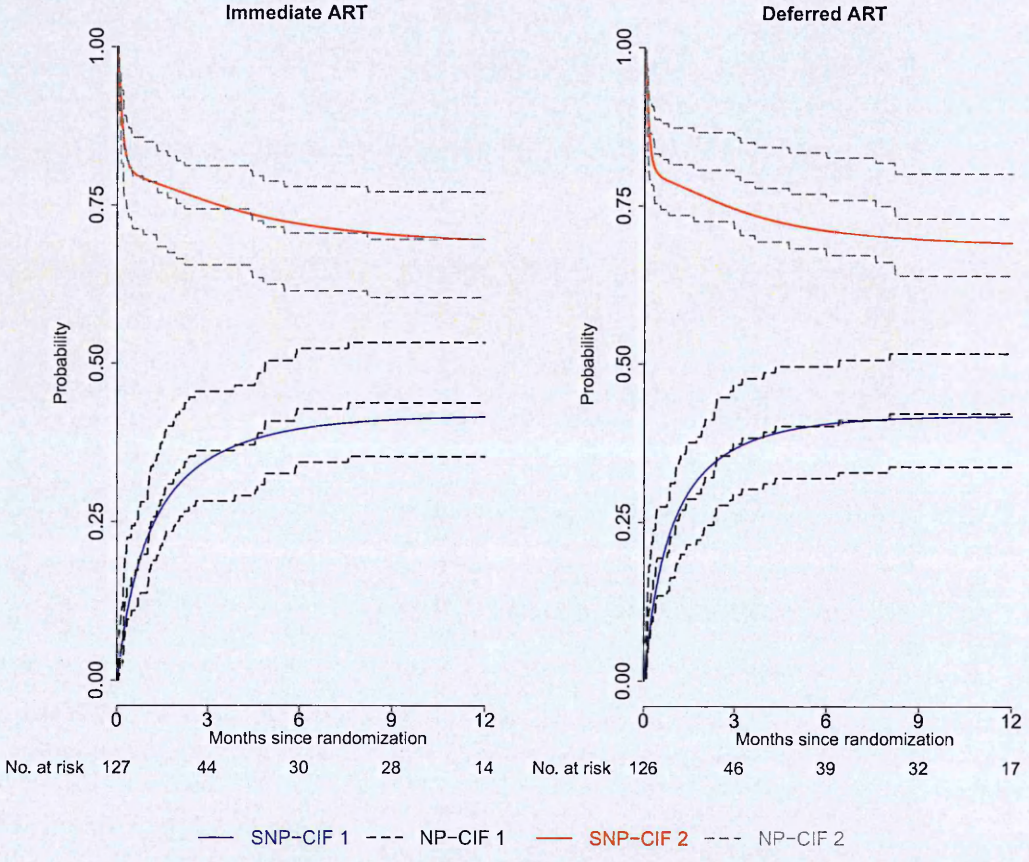
	Estimate[$]	"Asymptotic" SE	"Asymptotic" 95%-CI	Bootstrap SE*	Bootstrap 95%-CI*
Probability of $D = 2$ vs. $D = 1$					
- Immediate ART: yes	-0.20	0.29	(-0.77, 0.38)	0.33	(-0.91, 0.38)
Conditional time-to-event distribution for $T \mid D = 1$					
- Immediate ART: yes	0.23	0.27	(-0.31, 0.76)	0.29	(-0.42, 0.77)
Conditional time-to-event distribution for $T \mid D = 2$					
- Immediate ART: yes	0.21	0.22	(-0.21, 0.63)	0.34	(-0.52, 0.89)

SE: standard error, CI: confidence interval, $D = 1$: neurological event, $D = 2$: prior death.
[$] log-odds ratio for $D = 1$ in the first model, log-acceleration factors for the conditional time-to-event models. * based on 1000 bootstrap samples.

As shown in Figure 4.7.1, CIF estimates implied by the simple regression model and the nonpara-

metric estimates are relatively close to each other. An application of the informal test presented in Section 4.6.2 to assess whether the observed data is compatible with the parametric fit implied by the regression model also lead to large p-values ($p > 0.05$ for all 4 CIFs).

Figure 4.7.1: CIF for the time to neurological even and one minus CIF for time to prior death by treatment arm.



Event type 1: Neurological event. Event type 2: Prior death.
SNP-CIF: SNP estimate of the CIF based on a regression model including treatment only.
NP-CIF: nonparametric estimate of the CIF (with 95%-pointwise confidence intervals).

The baseline TBM grade is considered a strong predictor of outcome in patients with TBM with higher TBM grade associated with worse outcome (Török et al. (2011)). A regression model which included TBM grade in addition to treatment for all sub-models is summarized in Table 4.9. As for the simple regression model, the treatment assignment does not significantly affect any of the regression models after adjustment for TBM grade. This is not unexpected since randomized treatment assignment was stratified by baseline TBM grade.

Compared to baseline TBM grade I having TBM grade III was associated with a significantly lower marginal probability of experiencing a neurological event. This is not surprising as a higher TBM grade is known to be strongly associated with higher overall mortality leading to a higher chance to

die prior to manifestation of any neurological events. Moreover, a more severe TBM grades seems to accelerate the conditional time to both prior death and first neurological event with a more dramatic effect on the conditional time to prior death. However, these effects did not reach statistical significance according to bootstrap-based confidence intervals (which were considerably wider than their "asymptotic" counterparts).

Table 4.9: Estimates of regression coefficients and 95%-CIs from the multiple regression model.

	Estimate[$]	"Asymptotic" SE	"Asymptotic" 95%-CI	Bootstrap SE*	Bootstrap 95%-CI*
Probability of $D = 2$ vs. $D = 1$					
- Immediate ART: yes	-0.27	0.31	(-0.87, 0.34)	0.38	(-1.11, 0.36)
- TBM grade 2 (vs. 1)	-0.82	0.45	(-1.69, 0.05)	0.66	(-2.30, 0.23)
- TBM grade 3 (vs. 1)	-1.61	0.45	(-2.50, -0.73)	0.93	(-3.76, -0.60)
Conditional time-to-event distribution for $T \mid D = 1$					
- Immediate ART: yes	0.29	0.30	(-0.30, 0.87)	0.50	(-0.56, 1.38)
- TBM grade 2 (vs. 1)	-0.37	0.37	(-1.09, 0.35)	0.61	(-1.12, 1.32)
- TBM grade 3 (vs. 1)	-0.22	0.39	(-0.99, 0.54)	0.66	(-1.33, 1.30)
Conditional time-to-event distribution for $T \mid D = 2$					
- Immediate ART: yes	0.16	0.14	(-0.11, 0.43)	0.44	(-0.62, 1.14)
- TBM grade 2 (vs. 1)	-1.08	0.20	(-1.47, -0.69)	1.33	(-2.96, 1.78)
- TBM grade 3 (vs. 1)	-1.42	0.25	(-1.91, -0.93)	1.81	(-2.77, 3.72)

SE: standard error, CI: confidence interval, $D = 1$: neurological event, $D = 2$: prior death.
[$] log-odds ratio for $D = 1$ in the first model, log-acceleration factors for the conditional time-to-event models. * based on 1000 bootstrap samples.

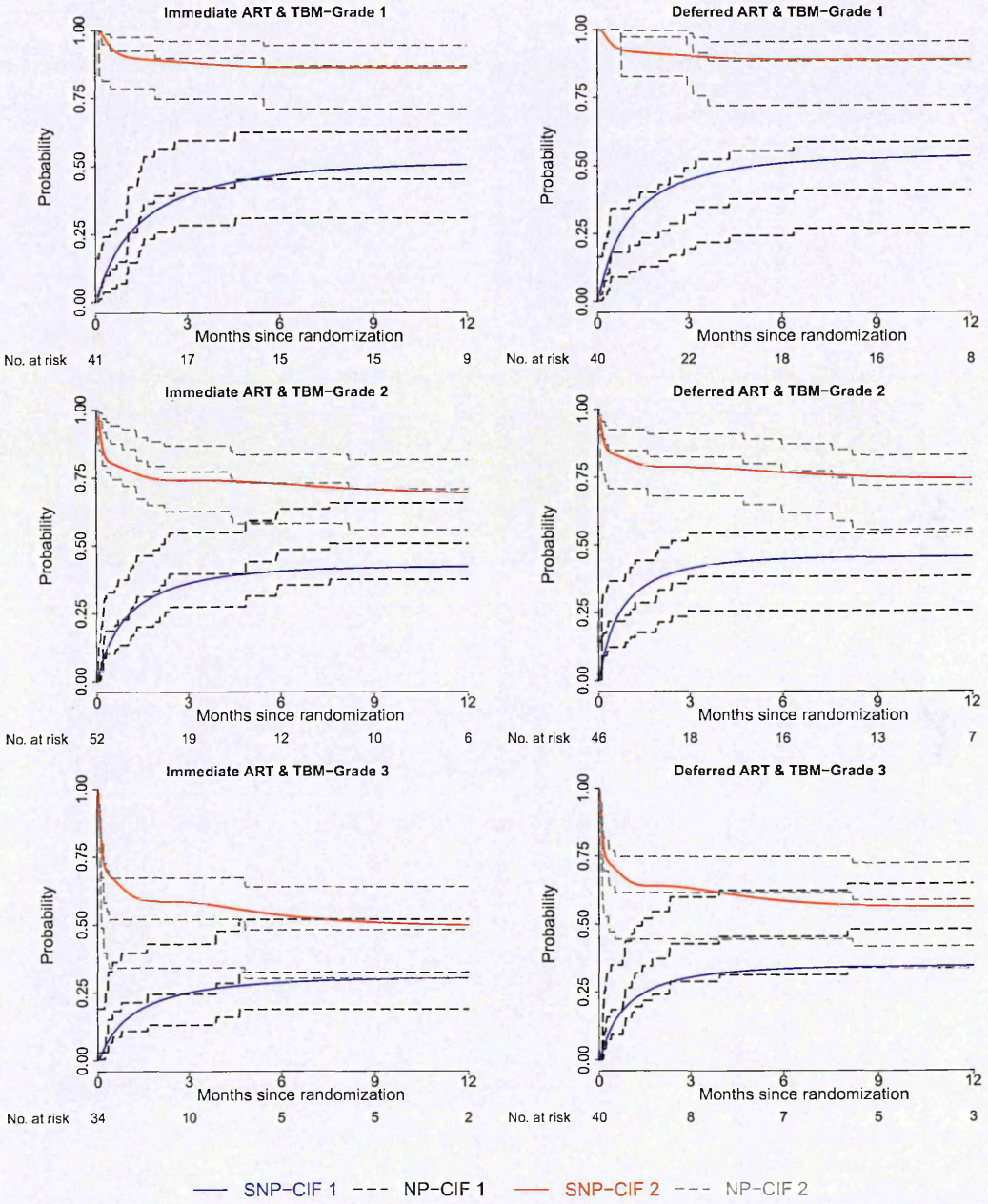
Figure 4.7.2 compares the fitted SNP CIF for each stratum specified by treatment arm and TBM grade. This shows reasonable agreement between SNP and nonparametric CIF estimates with the former almost always within the 95%-CI of the nonparametric estimates. One exception is the CIF of the time to death before experiencing any neurological event in TBM patients with grade 1 assigned to deferred ART in the early period where no neurological event was observed (though this could also be a problem of the 95% CI for the nonparametric CIF estimate which is unreliable at those time points). This is also the only instance where the $C_{t_{\max}}$ test (Section 4.6.2) showed a significant result (p-value = 0.017).

Several of the diagnostic tools presented in Section 4.6 were also used to verify the final SNP fit by regarding it as a parametric model. To apply them, I first generated 9 imputed data sets where event types for right-censored observations were imputed according to the method discussed in Section 4.6.3. For each of these data sets, plots of Pearson residuals (for the multinomial logistic model) and Cox-Snell and deviance residuals for each of the AFT models are displayed in Figures 4.7.3, 4.7.4 and 4.7.5, respectively.

Cox-Snell and deviance residuals plots suggest that the final SNP model fits well across imputed

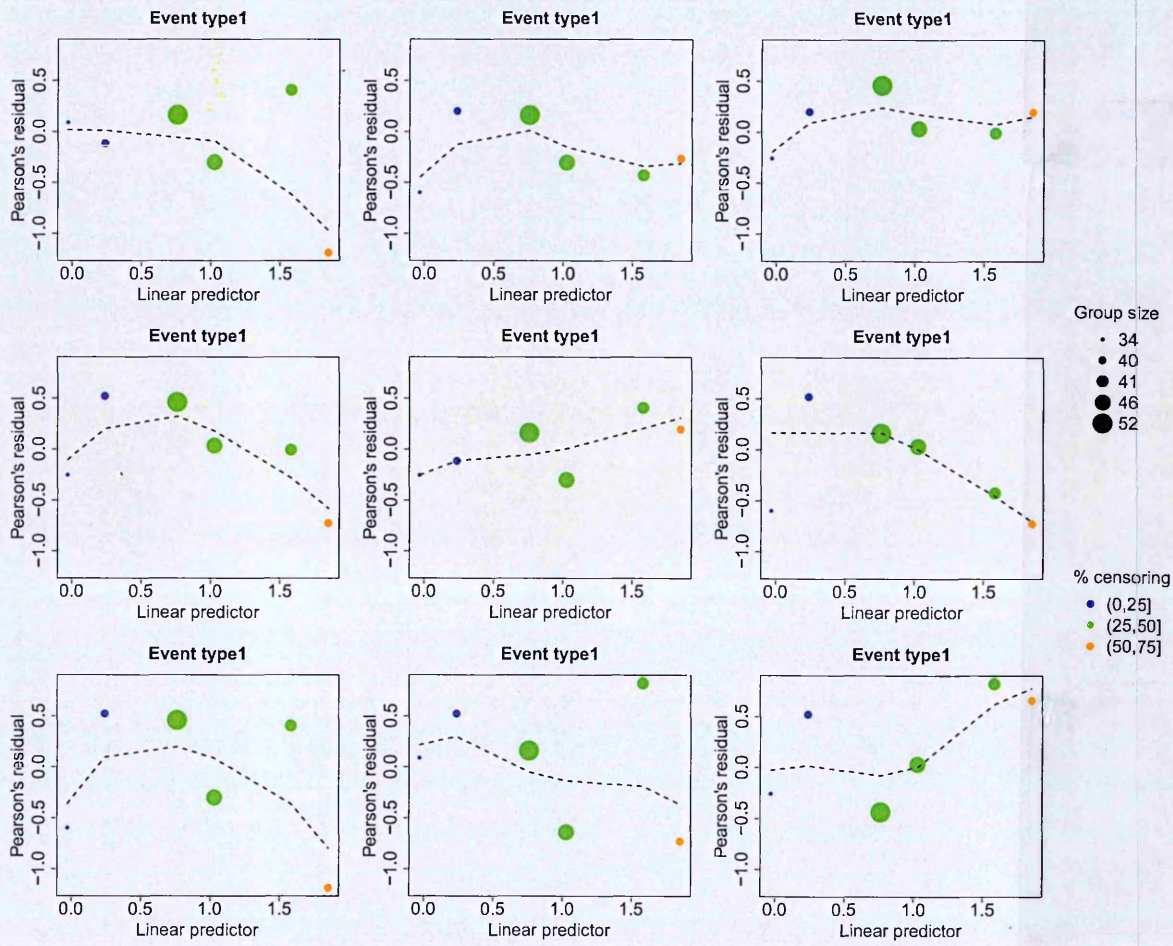
data sets. Pearson residuals for the logistic regression model were substantially affected by the by imputation; especially for strata of small size or with high censoring. However, all Pearson residuals were small in absolute value. Finally, I calculated martingale residual as defined in Section 4.6.4 for each event type based on the respective cumulative cause-specific hazards implied by the SNP model and displayed them in Figure 4.7.6. These also show no noticeable evidence against the final SNP fit. In conclusion, according to the aforementioned diagnostic results, the fitted SNP model, once regarded as a parametric model, well-describes the data.

Figure 4.7.2: CIF for the time to neurological event and one minus CIF for prior death by treatment arm and TBM-grade.



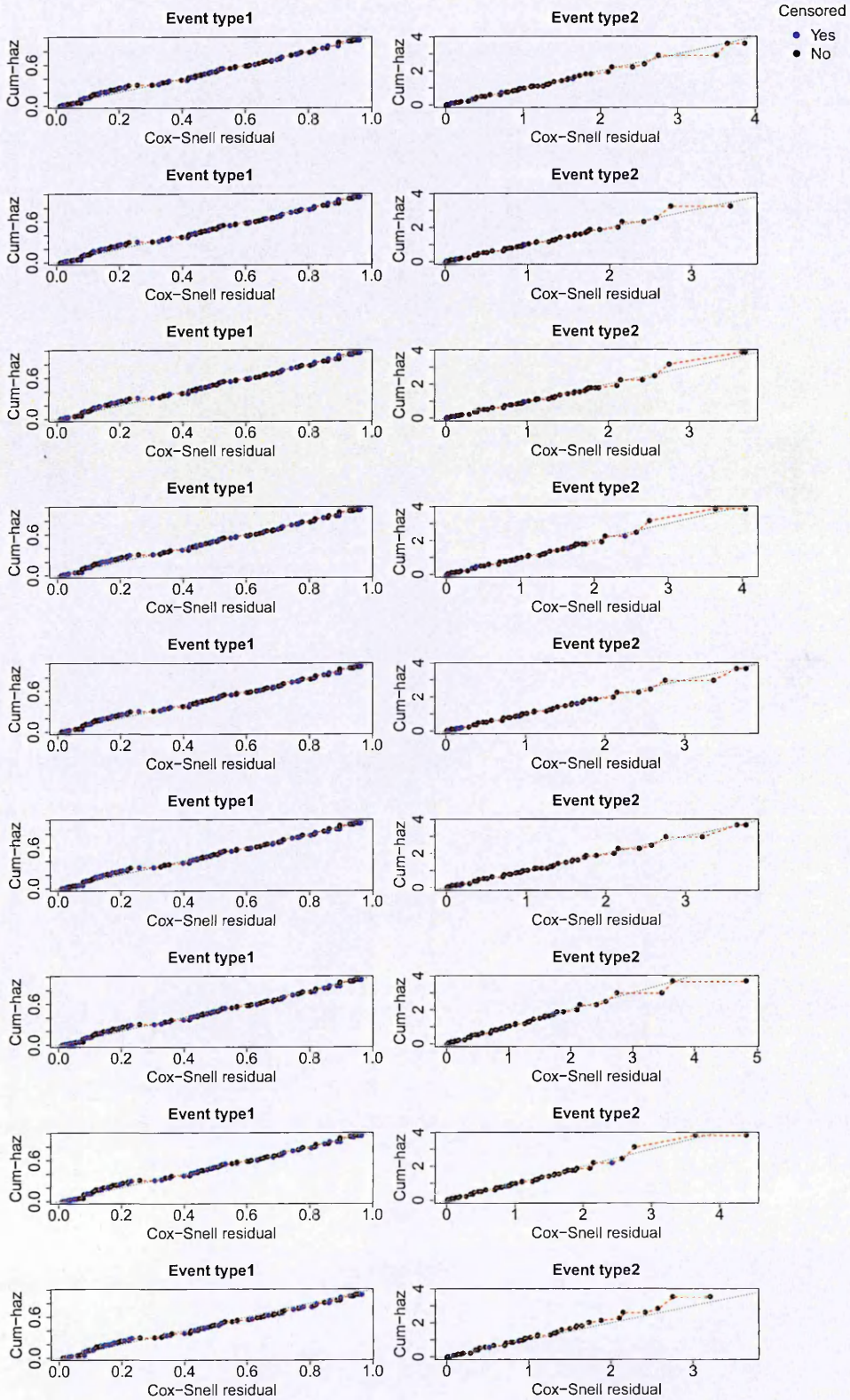
Event type 1: Neurological event. Event type 2: Prior death.
SNP-CIF: SNP estimate of the CIF based on a regression model including treatment and TBM grade.
NP-CIF: nonparametric estimate of the CIF (with pointwise 95% confidence intervals).

Figure 4.7.3: Pearson residual for 9 imputed data sets for the multinomial logistic sub-model of the SNP competing risks model.



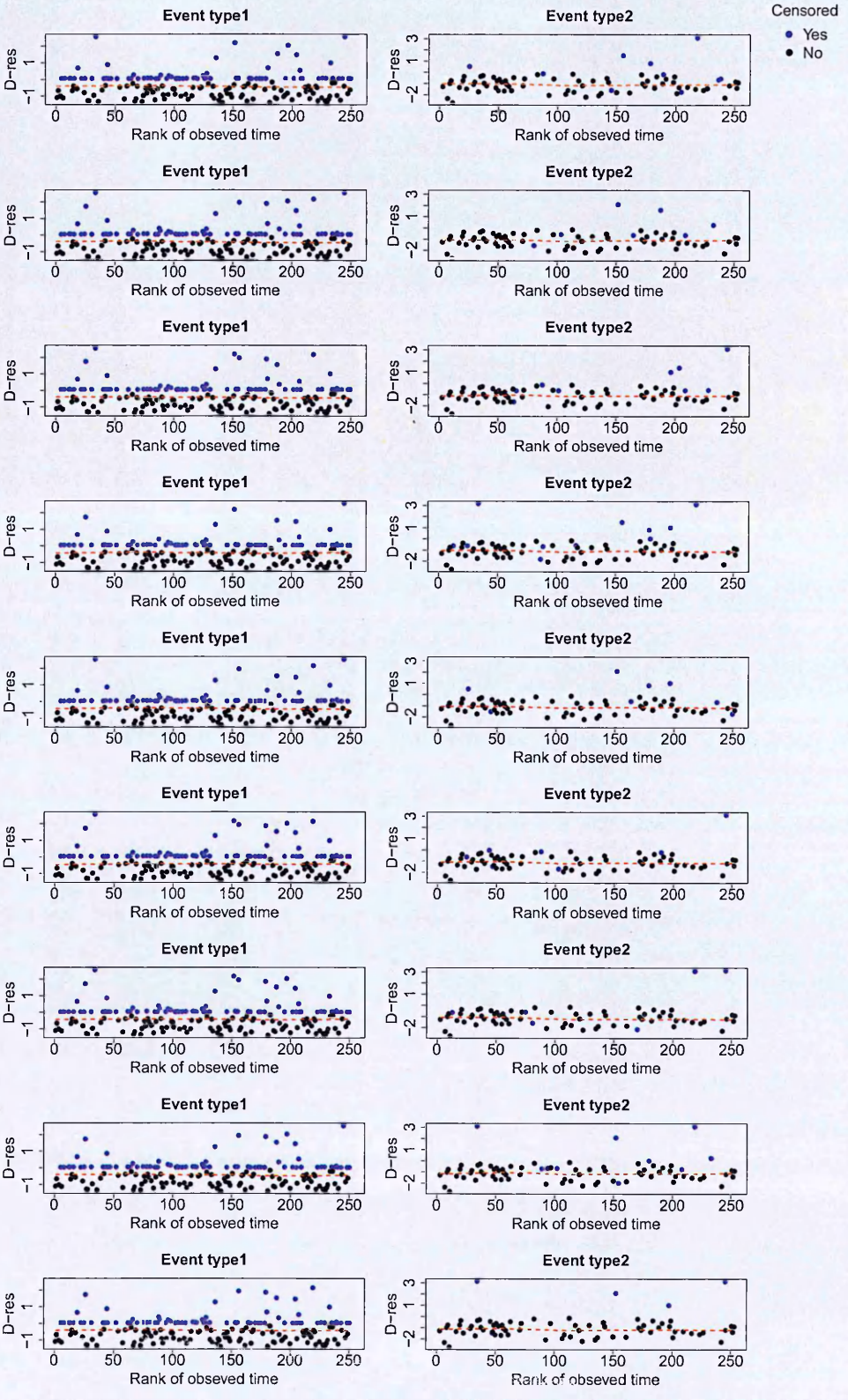
As there are only two event types, the multinomial logistic sub-model reduces to binary logistic regression and Pearson residuals for the occurrence of event type 1 (neurological event) are shown.

Figure 4.7.4: Cumulative hazard plots for 9 imputed data sets for the AFT sub-models of the SNP competing risks model.



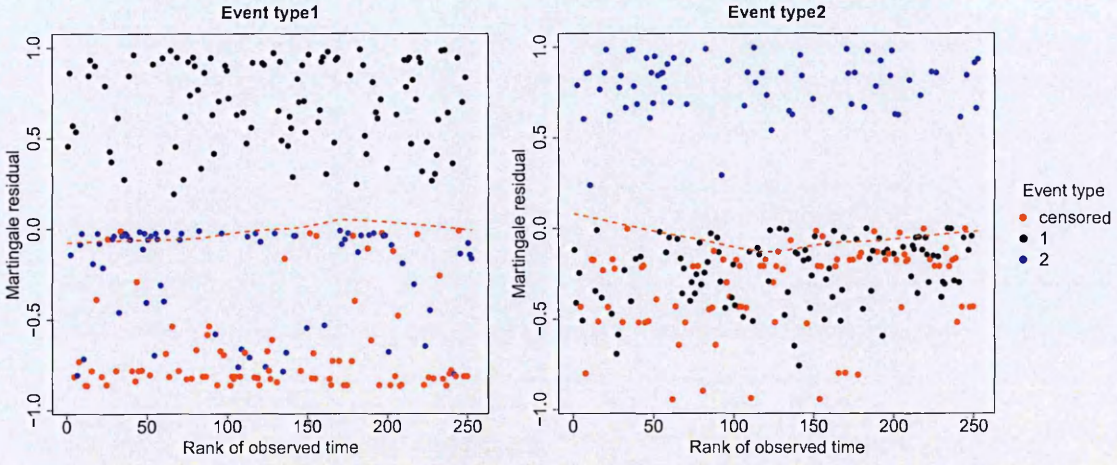
Event type: 1-Neurological event, 2-Prior death. Cum-haz: cumulative hazard.

Figure 4.7.5: Deviance residual for 9 imputed data sets for the AFT sub-models of the SNP competing risks model.



Event type: 1-Neurological event, 2-Prior death. D-res: Deviance residual.

Figure 4.7.6: Martingale residual for the whole competing risks process (based on the cumulative cause-specific hazards implied by the SNP competing risks model).



Event type: 1-Neurological event, 2-Prior death.

4.7.2 Combination antifungal therapy for cryptococcal meningitis

Here, I revisit the data set introduced in Subsection 3.7.2 which describes the interval-censored time to fungal clearance (beneficial event) or death prior to fungal clearance (competing event) in HIV-positive patients with cryptococcal meningitis. The main covariate is the treatment assignment (combination therapy of Flucytosine plus Amphotericin B versus Amphotericin B monotherapy). Other covariates of interest are the binary covariate indicating whether the baseline Glasgow comma score (GCS) was smaller than 15 and the continuous covariate of baseline fungal count (log10-transformed). According to the clinical publication of the trial Day et al. (2013), both an impaired level of consciousness (GCS<15) and a higher baseline fungal count were independent predictors of 6-month mortality. For simplicity, I included only observations with non-missing values for all covariates in the regression models which comprise of 68/86 patients in the combination therapy group and 74/89 patients in the monotherapy group. The missing data was almost entirely due to missing baseline fungal counts and, due to the relatively large proportion of missing data, a more thorough analysis might be based on multiple imputation.

Results of the simple regression model are reported in Table 4.10, which shows that the combination therapy significantly increases the marginal probability of reaching fungal clearance. This is in agreement with the results in Subsection 3.7.2 as well as the findings reported in Table 2 of Day et al. (2013). It is less clear whether or not treatment really affects the conditional time to fungal clearance, and neither "asymptotic" nor bootstrap-based 95%-CIs showed a significant treatment effect on the conditional time to prior death. Moreover, it should be noted again that treatment effects for the

conditional time to event distributions should be interpreted with caution because by conditioning on the event type they do not represent causal effects even in a randomized trial.

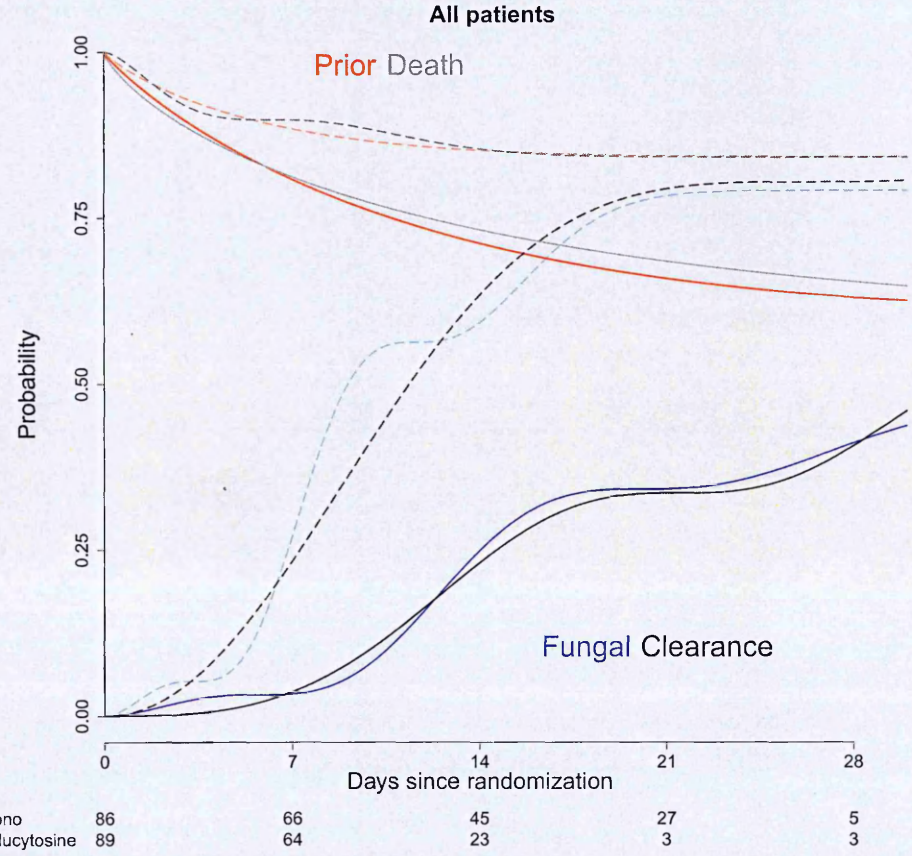
Table 4.10: Results from the simple regression model.

	Estimate[$]	"Asymptotic" SE	"Asymptotic" 95%-CI	Bootstrap SE*	Bootstrap 95%-CI*
Probability of $D = 2$ vs. $D = 1$					
- Combination therapy: yes	1.41	0.43	(0.57, 2.25)	0.49	(0.48, 2.43)
Conditional time-to-event distribution for $T \mid D = 1$					
- Combination therapy: yes	-0.55	0.07	(-0.69, -0.40)	0.31	(-0.90, 0.32)
Conditional time-to-event distribution for $T \mid D = 2$					
- Combination therapy: yes	-0.69	0.46	(-1.59, 0.20)	0.68	(-2.31, 0.49)

SE: standard error, CI: confidence interval, $D = 1$: fungal clearance, $D = 2$: prior death.
[$] log-odds ratio for $D = 1$ in the first model, log-acceleration factors for the conditional time-to-event models. * based on 1000 bootstrap samples.

In Figure 4.7.7 I visually compared the estimated CIFs from separate SNP estimation (Subsection 3.7.2) and the simple regression model. Unlike the previous example, there is a bigger disparity between the different SNP methods. In fact, the SNP fit of the simple regression model has $K_1 = 2$ and $K_2 = 1$, whereas the SNP models for the combination therapy group and the monotherapy group have fitted SNP distributions with, respectively $(K_1 = 1,\ K_2 = 1)$ and $(K_1 = 1,\ K_2 = 0)$. All SNP models selected exponential base densities.

Figure 4.7.7: CIF for the time to fungal clearance (blue for the simple regression model and black for separate CIF estimation) and one minus CIF for prior death (red for the simple regression model and gray for separate CIF estimation) by treatment arm.



Solid lines correspond to Amphotericin B monotherapy, dashed lines to Amphotericin B + Flucytosine combination therapy.

Multiple regression results are displayed in Table 4.11: Estimates for the treatment effect were roughly comparable to the unadjusted analysis. Estimates for the other covariates showed the clinically expected direction (a higher marginal risks of prior death, slower conditional time to fungal clearance and faster conditional time to death for impaired consciousness and higher baseline fungal counts) but none of these effects except for GCS<15 for the conditional time to fungal clearance reached statistical significance according to the bootstrap confidence intervals. As in the previous example, all bootstrap-based confidence intervals were much wider than their "asymptotic" counterparts.

Table 4.11: Results from the multiple regression model.

	Estimate[$]	"Asymptotic" SE	"Asymptotic" 95%-CI	Bootstrap SE*	Bootstrap 95%-CI*
Probability of $D = 2$ vs. $D = 1$					
- Combination therapy: yes	1.82	0.49	(0.85, 2.79)	1.46	(-0.10, 3.53)
- GCS < 15: yes	-1.41	0.53	(-2.45, -0.36)	1.50	(-2.97, 0.71)
- log10 baseline fungal count	-0.46	0.24	(-0.93, 0.00)	0.33	(-0.90, 0.41)
Conditional time-to-event distribution for $T \mid D = 1$					
- Combination therapy: yes	-0.39	0.08	(-0.55, -0.22)	0.24	(-0.62, 0.43)
- GCS < 15: yes	0.38	0.10	(0.18, 0.58)	0.25	(0.07, 1.01)
- log10 baseline fungal count	0.17	0.04	(0.10, 0.24)	0.11	(-0.14, 0.24)
Conditional time-to-event distribution for $T \mid D = 2$					
- Combination therapy: yes	-1.05	0.48	(-1.99, -0.10)	1.14	(-3.60, 0.85)
- GCS < 15: yes	-0.14	0.46	(-1.04, 0.76)	1.12	(-1.87, 1.87)
- log10 baseline fungal count	-0.85	0.30	(-1.43, -0.26)	0.55	(-1.94, 0.11)

SE: standard error, CI: confidence interval, $D = 1$: fungal clearance, $D = 2$: prior death.
[$] log-odds ratio for $D = 1$ in the first model, log-acceleration factors for the conditional time-to-event models. * based on 1000 bootstrap samples.

4.8 Discussion

This chapter first introduced a competing risks regression model by including covariates into the marginal and conditional models of the SNP method for CIF estimation proposed in Chapter 3. This offers a flexible way to describe covariate effects on the CIFs. The simulation studies demonstrated that the proposed model can lead to substantially more precise parameter estimates (compared to parametric models) and predictions (compared to parametric and semi-parametric models) when the comparator models were not correctly specified. In contrast, estimates and predictions were only modestly improved by parametric and semiparametric models compared to my approach if data was simulated according to those alternative models. Moreover, to my knowledge, this is the only available flexible model for competing risks regression modelling in the presence of interval censoring where currently only parametric models have been proposed.

Despite these strengths, my SNP model carries the intrinsic weaknesses of a model based on the mixture factorization (4.1.1). First, the marginal component $P(D)$ is poorly identified if the data includes insufficient follow-up relative to the timing of events. The simulations in Chapter 3 showed that this identifiability issue does not seem to severely affect CIF estimation within the observed follow-up period. However, the problem is more severe here as one cannot restrict attention to the observed follow-up period because the regression coefficients on the marginal component model $P(D)$ describe the ultimate event state at time infinity. A practical implication is that the SNP model is primarily useful if $P(D)$ is reasonably well identified from the data. The second limitation of using the mixture factorization is that it conditions on the future event status D which makes interpretation of regres-

sion coefficients difficult, especially for covariates affecting the conditional components. However, if regression coefficients are only allowed to affect the marginal component, interpretation is much easier because, as discussed in Section 4.3, covariate effects on the marginal component can directly be translated to covariate effects on the CIFs at any time point. Of note, parametric competing risks models (Larson & Dinse (1985), Lau et al. (2008) and Lau et al. (2011)) as well as cure rate survival models (Kuk & Chen (1992)) based on the mixture factorization also suffer from the problems discussed in this paragraph but these limitations are often not adequately discussed in the respective literature.

Another limitation of the SNP model relates to its "ad hoc" asymptotic statistical inference which ignores the adaptive choice of the polynomial coefficients. Simulations in Chapter 3 showed that this usually leads to reasonable inference for the purpose of CIF estimation. In the present regression setting, coverage of "asymptotic" confidence intervals was frequently close to nominal levels for a sample size of $n = 500$ (though there was some indication of undercoverage) but for the lower sample size of $n = 100$, undercoverage was substantial. In line with this, bootstrap standard errors and confidence intervals in the applications were much larger than the "ad hoc" asymptotic counterparts. While the validity of bootstrap-based statistical inference itself would be difficult to justify mathematically, it is nevertheless expected that bootstrap-based confidence intervals are more reliable. This has also been shown in simulations for SNP survival models (Zhang & Davidian (2008)) but could not be investigated here as it was computationally infeasible.

The second theoretical part of this chapter gave a brief overview of diagnostic tools for right-censored competing risks regression models. Literature on diagnostics for competing risks methods is largely lacking and the proposed methods can be used for the final SNP fit of my model as well as for several parametric models.

In conclusion, the SNP regression model presented in this chapter is a an extension of the model in Chapter 3. It provides a flexible way to model covariate effects on the CIF under arbitrary censoring. The model performed well in simulation studies in terms of accuracy of parameter estimates and predictive ability. However, several limitations related to identifiability, interpretation, and validity of asymptotic inference remain.

Chapter 5

Weighted analyses of composite endpoints

The previous chapters focused on methods for competing risks where the time and type of the first disease event is of interest. This chapter focuses on composite endpoints which are increasingly used as primary endpoints in randomized controlled clinical trials (RCTs). Competing risks methods can be used to analyse composite endpoints but, as will be discussed, more general methods which also consider disease events occurring after the first event might be more relevant. Specifically, this chapter suggests novel test statistics for the weighted comparison of composite endpoints. Weights are introduced to address one of the shortcomings of conventional analyses of composite endpoints which ignore that different disease events included in the composite endpoint might differ in their clinical importance.

This chapter is structured as follows: Section 5.1 gives a short overview of the composite endpoints literature. Section 5.2 introduces a general framework for the weighted analysis of composite endpoint considering both binary and time to event combined endpoints. Section 5.3 proposes a strategy for simultaneous inference across multiple weighting schemes. A simulation study which investigates the actual performance of the proposed multiplicity adjustment is given in Section 5.4, followed by applications (Section 5.5) and concluding remarks (Section 5.6).

5.1 Composite endpoints, a short overview

5.1.1 Composite endpoints in clinical studies

Traditionally a composite endpoint is defined as the occurrence of at least one of a given set of different disease events (also denoted as "component outcomes") within a certain follow-up period (Ferreira-González et al. (2007)). For example, in cardiology studies, the component outcomes could be myocardial infarction, stroke or death, and the conventional analysis would either focus on the

binary outcome whether any of these events occurred or on the time to the first event. Composite endpoints have been widely used in a large spectrum of clinical disciplines including oncology (Mell & Jeong (2010)) and cardiology (Armstrong et al. (2011)). For example, Lim et al. (2008) showed that 37% of RCTs in cardiovascular medicine and surgery published between 2000 and 2006 reported a composite endpoint with a median of three individual component outcomes. Composite endpoints have also been used in RCTs of infectious diseases conducted by OUCRU-VN. Our trials in typhoid fever have used the composite primary endpoint of treatment failure defined as the occurrence of at least one of the following component outcomes: prolonged fever clearance time, need for rescue treatment, microbiological failure, relapse, or enteric-fever-related complications (Pandit et al. (2007) and Arjyal et al. (2011)). For trials in CNS infections such as tuberculous and cryptococcal meningities, we have reported the composite endpoints of the time to the first neurological event or death and the occurrence of disability or death at the end of follow-up (Thwaites et al. (2004), Török et al. (2011) and Day et al. (2013)).

The widespread use of composite endpoints can be explained by the following clinical and statistical benefits. First, evaluating an intervention often requires looking at different relevant clinical outcomes which can represent various pathophysiological aspects of the disease process of interest and the composite endpoints provides an overall summary measure of the impact of an intervention (Cannon (1997), Freemantle & Calvert (2007*b*), Sampson et al. (2010) and Tong et al. (2012)). Similarly, the combination of efficacy and safety properties of an intervention might be useful (Ferreira-González et al. (2007)). Second, from a statistical point of view, using composite endpoints can potentially increase the statistical power for detecting an effect of an intervention compared to focusing on a single component endpoint (Cannon (1997)). Indeed, as the incidence of the most severe outcomes including death has been reduced due to medical advances in many fields, RCTs with these most severe and often most clinically relevant outcomes as primary endpoints have become infeasible. Hence, the anticipated power gains (and corresponding decreases in the required sample size) from using a combined endpoint which includes death and less severe outcomes has probably been the main reason for the increasing usage of composite endpoints.

Despite these strengths, traditional approaches to analysing composite endpoints do have certain limitations. Most comparisons of composite endpoints between treatment groups only compare whether any of the component outcomes occur (binary outcome) or the time to first outcome (survival outcome). This pooling of component outcomes of potentially varying clinical importance may lead to misleading perception of treatment efficacy because the result is driven by component outcomes which are observed more frequently and earlier but are often less severe and of less clinical importance than the less frequent and later events. For example, in studies of acute coronary diseases one of

the less common outcomes is death, the most important/severe outcome (Armstrong et al. (2011)). Moreover, the effects of an intervention on different component outcomes (component effects) may differ in direction. This not only complicates the interpretation of the overall treatment effect, which is essential for choosing an optimal treatment, but may also remove or even reverse the power benefits of using composite endpoint (Ferreira-González et al. (2007)). Moreover, a systematic review of RCTs reporting binary composite endpoints concluded that component outcomes are often unreasonably combined, inconsistently defined, and inadequately reported (Cordoba et al. (2010)).

5.1.2 Statistical analysis of composite endpoints

As mentioned above, in most analyses, component outcomes are pooled to a single binary or time-to-event endpoint. Accordingly, standard statistical methods such as tests for the comparisons of proportions or logistic regression (for binary outcomes) and log-rank tests or the Cox proportional hazards model (for time-to-event outcomes) are usually used for the statistical comparison between the two groups. However, several alternative analyses methods have been suggested in the literature and some of them are outlined in this and the next subsection.

In the case of binary outcomes, several authors have suggested to analyse the occurrence of different component outcomes as a multivariate outcome and demonstrated that associated overall tests can be more powerful than the simple analysis of the collapsed outcome (Lefkopoulou & Ryan (1993) and Mascha & Imrey (2010)). A review article of analysis methods for binary composite endpoints which favours multivariate methods is Mascha & Sessler (2011). In their approach each subject carries multiple records representing their observed component outcomes. The treatment effect and the effects of other covariates on each component outcome are then modelled with a binary logistic regression model and correction for potential correlation between different patient records is done using generalised estimating equation, where the correlation between component outcomes from the same subject is modelled by an a priori specified within-subject correlation structure. Compared to the traditional approach, this method can produce results that are less driven by the more frequent components. In addition, it can incorporate various covariates and allows for testing heterogeneity of treatment effects across components. However, one disadvantage of their approach is that the analysis is more complex and hence more difficult to interpret for a clinical audience. Moreover, treatment effects on component outcomes might be difficult to interpret in the situation of competing risks, e.g. an intervention associated with a higher mortality might also be associated with a lower probability of less severe outcomes simply because the pool of survivors that can potentially experience this component is reduced.

If individual component outcomes of a composite outcome are analysed, this poses two additional

problems. First, RCTs with a composite primary outcome are often not adequately powered to detect treatment effects on component outcomes. Second, issues related to multiple testing arise. One method to deal with the latter is to use gatekeeping procedures which require organizing all hypotheses of interest into an ordered sequence of sets. For each set, significance tests which preserve the type I error within that set are defined. The full procedure than sequentially proceeds through the sets but is only allowed to move to the next set if the overall test on the previous set reaches statistical significance, i.e. the "gate to proceed is open". A review of gatekeeping procedures applied to composite endpoints is provided by Mascha & Turan (2012).

One method for analysing individual component outcomes for time-to-event outcomes is to focus on the time and type of the first event and then to apply competing risks methods, the topic of the previous chapters of this thesis. Such analyses suffer from the same problems as mentioned in the previous paragraph. One approach to address these issues has been proposed by Rauch & Beyersmann (2013): They suggest to first order the component outcomes based on their clinical importance. Then, in the simplest case considered, a trial should be powered to show not only superiority of the tested treatment on the composite endpoint but also non-inferiority with respect to the most important component outcomes. In this framework, superiority and non-inferiority are investigated by using the overall hazard of the composite endpoint and the cause-specific hazards of the component outcomes, respectively.

Moreover, competing risks analyses applied to composite endpoints have one additional shortcoming: By focusing on the first event, they neglect that the first event may be followed by a subsequent component outcome which is more severe and clinically relevant. Hence, competing risks analyses applied to composite endpoints are not always useful and may even be misleading (Wolbers et al. (2014)). To model first events and subsequent events jointly, more complex multistate models are required (Beyersmann et al. (2012)).

5.1.3 Weighted analyses

As discussed, one major problem of composite endpoints is that they weight all component endpoints equally whereas in reality, the clinical importance of component outcomes may differ substantially. Hence, several authors stressed the importance of weighting the component to improve the clinical relevance and interpretation of composite endpoint analyses (e.g. Ferreira-González et al. (2007), Hong et al. (2011) and Tong et al. (2012)).

Several approaches to weighted analyses of composite endpoints have been proposed. For binary outcomes, Sampson et al. (2010) proposed assigning a weight or score to each component outcome, then summing up the scores corresponding to the component outcomes for each patient and com-

paring the total patient scores between treatment arms based on the Mann-Whitney U test. Weights can also easily be added to the multivariate analyses for composite binary outcomes outlined above (Mascha & Sessler (2011)). For time-to-event outcomes, Bakal et al. (2012) suggested a weight adjustment to the Kaplan-Meier estimator. In this approach, all weights are constrained to be ≤ 1 and full weights of 1 affect the Kaplan-Meier estimator as usual whereas fractional weights only remove that fraction of the subject from the risk set. A disadvantage of all of these methods is that the resulting test statistics are somewhat difficult to interpret. To circumvent this, I will present an alternative test statistic with a straightforward clinical interpretation in Section 5.2.

Assigning exact quantitative weights to different component outcomes is challenging. First, in many clinical settings, elucidation of quantitative weights faces dissent between experts. Second, even when experts can reach consensus, individual patients might assess component endpoints differently (Glasziou et al. (1990) and Freemantle & Calvert (2007*a*)). Nevertheless, several proposals have been made in the literature. Recent proposals include deriving weights based on a clinician-investigator Delphi panel (Armstrong et al. (2011)), discrete choice experiment amongst patients (Tong et al. (2012)), or disability-adjusted life years lost (Hong et al. (2011)). Moreover, ignoring weights altogether leads to the conventional analysis of composite endpoints which implicitly assigns equal weight to all first observed component outcomes and ignores subsequent outcomes altogether which appears to be even less meaningful. While assigning exact weights might be difficult in many settings, it is frequently possible to rank component outcomes according to their relative importance. To exploit this, I will introduce multiplicity adjustments for the proposed test statistic which either allow for simultaneous inference across all sets of weights, all weights following an ordering constraint, or more general constraints, respectively (see Section 5.3).

5.2 A unified framework for weighted analyses of binary and time-to-event composite endpoints

5.2.1 Notation and proposed test statistics

The previous section laid down several benefits of weighted analyses of composite endpoints. As seen from this short review, several approaches to weighted analyses of composite endpoints have been proposed but interpretation of the suggested test statistics is not straightforward. In this section, I propose a unified framework for weighted analyses of composite endpoints that can be used in the context of randomized clinical trials for both binary and time to event data.

Specifically, I assume that interest is in the occurrence of certain clinical event types $k = 1, \ldots, K$ during a predefined follow-up period $(0, \tau]$. For a given weight vector $w = (w_1, \ldots, w_K)^T \in \mathbb{R}^K$,

I propose the following weighted test statistic for comparing the respective event type proportions between two independent groups A and B

$$\mathcal{T}(w,\tau) = \sum_{k=1}^{K} w_k \left(\hat{p}_{A,k}(\tau) - \hat{p}_{B,k}(\tau)\right) = w^T \hat{\mathcal{D}}(\tau) \tag{5.2.1}$$

where $\hat{p}_{A,k}(\tau)$ and $\hat{p}_{B,k}(\tau)$ are the estimated absolute risks or probabilities that an event type k in group A and B occurs in the interval $(0,\tau]$, respectively. For convenience, from now on these quantities are referred to as event type probabilities. Using $\mathcal{T}(w,\tau)$ one can then conduct a Wald-type significance test of the two- (resp. one-) sided null hypothesis $H_0 : \mathcal{T}(w,\tau) = 0$ (resp. $H_0 : \mathcal{T}(w,\tau) \leq 0$) against the two- (resp. one-) sided alternative hypothesis $H_A : \mathcal{T}(w,\tau) \neq 0$ (resp. $H_A : \mathcal{T}(w,\tau) > 0$). Performing these tests is possible as long as $\hat{\mathcal{D}}(\tau)$ (and consequently $\mathcal{T}(w,\tau)$) follows an asymptotic normal distribution and a corresponding sample-based estimator of its covariance matrix is available. Estimation of these quantities for specific situations shall be discussed later in Subsection 5.2.3. Meanwhile, it is assumed that these quantities are obtainable.

Importantly, the proposed test statistic has a straightforward and clinically relevant interpretation. If the weights are standardized to sum to one, $\mathcal{T}(w,\tau)$ is the weighted average of absolute risk differences for individual event types. Moreover, if each weight represents a certain type of "cost" associated with event type k then $\mathcal{T}(w,\tau)$ estimates the expected cost difference between the two interventions A and B. Of note, there is no technical barrier preventing the use of positive weights for some event types and negative weights for others. However, as negative weights are not sensible for component events of a composite endpoint which are usually all harmful, only nonnegative weights are considered for simplicity in this chapter. However, weights of opposing signs might be useful to measure more general trade-offs between beneficial and harmful event types.

5.2.2 Event type definition

If the number of component outcomes of a composite endpoint is large and subjects can experience more than one component outcome, subjects may experience many different possible combinations of component outcomes over time. In principle, each unique combinations can constitute an event type in the terminology of the previous subsection ("exhaustive" setting) but there are also simpler settings which require the assignment of weights to a more manageable number of event types. In particular two alternative settings, the "competing risks" and the "marginal" setting will be described.

To simplify the discussion I assume that the data follows an illness-death model with a composite endpoint consisting of one fatal component outcome and one nonfatal component outcome. Moreover, to be more illustrative, let the nonfatal outcome be nonfatal myocardial infarction (MI) and let the

fatal one be death (DE) with or without a prior MI. In this case, the most "exhaustive" set of binary event types stemming from MI and DE is:

- MI^+DE^-: having MI and staying alive until τ, with weight $w_{MI^+DE^-}$.
- MI^-DE^+: dying in $(0, \tau]$ without any prior MI, with weight $w_{MI^-DE^+}$.
- MI^+DE^+: having MI and dying in $(0, \tau]$, with weight $w_{MI^+DE^+}$.
- MI^-DE^-: staying alive and MI-free until τ, with weight $w_{MI^-DE^-}$.

I refer to these event types as "exhaustive" event types and the whole setting is called the "exhaustive" setting. This definition of event types partitions the set of all subjects into four mutually exclusive event type categories.

In general, it makes sense to give higher weights to event types that are clinically more severe or important. Under the current "exhaustive" setting this implies $w_{MI^+DE^+} \geq w_{MI^-DE^+} \geq w_{MI^+DE^-} \geq w_{MI^-DE^-}$. One may also make allowances for exact zero weight(s) e.g. letting $w_{MI^+DE^-} = 0$ if for some reason MI^+DE^- were not of interest. Moreover, it is usually sensible to exclude the event type MI^-DE^- (no event) from the test statistic altogether (which is equivalent to setting $w_{MI^-DE^-} = 0$). This also has the benefit that the resulting covariance matrix of the test statistic $\hat{\mathcal{D}}(\tau)$ is of full rank (whereas it would be singular otherwise) which is required for the multiplicity adjustment method presented in Section 5.3.

A limitation of the "exhaustive" setting is that the number of "exhaustive" event types may grow very fast as the number of binary component outcomes increases, e.g. exponential growth when all component outcomes are nonfatal. Such a growth in the number of event types may lead to difficulties in choosing weights for all possible combinations and sparsity in the respective observed numbers of subjects which may make the variance estimator for the distribution of the respective $\mathcal{T}(w, \tau)$ statistic unstable. An alternative setting with fewer event types is as follows:

- MI^+: experiencing MI as a first event (with or without later death) in $(0, \tau]$, with weight w_{MI^+}.
- MI^-DE^+: experiencing death in $(0, \tau]$ without a prior MI, with weight $w_{MI^-DE^+}$.
- MI^-DE^-: staying alive and MI-free until τ, with weight $w_{MI^-DE^-}$.

I call this setting "competing risks" setting because only the first observed event is relevant. As before, MI^-DE^- should be excluded in the calculation of the test statistics. Similar to the "exhaustive" setting, all event types considered here are mutually exclusive. Of note, this setting is a special case of the "exhaustive" setting with the additional weight constraint $w_{MI^+} = w_{MI^+DE^-} = w_{MI^+DE^+}$.

This weight constraint which represents ignorance of the test statistic with respect to what happens after the first MI is difficult to justify clinically. Hence, the "competing risks" setting is not generally recommended and not pursued further.

The final setting that I propose is the "marginal" setting which consists of the following event types:

- MI^+: MI with or without later death in $(0, \tau]$, with weight w_{MI^+}.
- DE^+: death with or without prior MI in $(0, \tau]$, with weight w_{DE^+}.
- MI^-DE^-: staying alive and MI-free until τ, with weight $w_{MI^-DE^-}$.

As before, the state MI^-DE^- is usually ignored and a natural weight constraint in this setting is $w_{DE^+} \geq w_{MI^+}$.

The major difference between this setting and the previous two is that it contains overlapping event types, i.e. subjects experiencing both MI and death experience both event types DE^+ and MI^+. Thus, it is theoretically possible that the respective observed proportions of event types MI^+ and DE^+ (in each group) sum up to more than one.

The marginal setting is also a special case of the exhaustive setting with the additional weight constraint that $w_{MI^+DE^+} = w_{DE^+} + w_{MI^+}$. This additional constraint which says that the "costs" of different component events sum up in an additive way appears reasonable in many contexts. For example in the current illness-death example, this setting appropriately accounts for the fact that MI^+DE^+ is clinically more severe than MI^+DE^-. Thus, the "marginal" setting achieves the same reduction in dimensionality as the "competing risks" setting but is often interpretationwise preferable.

As the event types of the "marginal" setting can easily be derived from the exhaustive event types by simply pooling them, the test statistic (and associated covariance matrix) $\hat{\mathcal{D}}(\tau)$ for this setting can be obtained as a simple linear transformation from the respective test statistics for the "exhaustive" events as discussed in the next subsection.

Of note, while the discussion in this subsection was for a simple illness-death model for illustrative purposes, the definition of "exhaustive", "competing risks" and "marginal" settings can be transferred to more complicated settings in a straightforward way.

5.2.3 Distribution of the proposed test statistic

In this subsection the distributional properties of the test statistic $\mathcal{T}(w, \tau)$ in Equation (5.2.1) are discussed for the different event type settings mentioned above. As before, this subsection also uses the notation given earlier in Subsection 5.2.1. For ease of discussion, I also keep using the myocardial

infarction (MI) and death (DE) example. However, the described theory can be extended to more general settings in a straightforward way.

The simplest scenario under which $\hat{\mathcal{D}}(\tau)$ (hence $\mathcal{T}(w,\tau)$) follows an asymptotic normal distribution is when all subjects have complete follow-up until time τ. Whence under the "exhaustive" setting, the associated event type probabilities (in each group) can be consistently estimated by the corresponding observed proportions and these correspond to the cell probabilities of a multinomial distribution. Hence, a covariance estimator for the respective vectors of proportions is directly available by plugging the observed proportions into the corresponding covariance matrix derived from this multinomial distribution. $\hat{\mathcal{D}}(\tau)$ can then be obtained as the difference of the two vectors of proportions for groups A and B and the associated covariance matrix is the sum of the two group covariance matrices for each group. Moreover, it follows from the central limit theorem result that $\hat{\mathcal{D}}(\tau)$ based on the "exhaustive" event types (excluding the "no event" category MI^-DE^-) and an associated weight vector $w_e = (w_{MI^+DE^-}, w_{MI^-DE^+}, w_{MI^+DE^+})^T \in \mathbb{R}^3_+$ has an asymptotic multivariate normal distribution with a consistent covariance estimator $\hat{V}_e$. Hence the estimated variance of $\mathcal{T}(w,\tau)$ is $w^T\hat{V}_e w$.

Under no right-censoring, as mentioned earlier the "marginal" event type proportions in group A can easily be calculated from the "exhaustive" ones by using an appropriate linear transformation. For the current example this linear relation is

$$\begin{pmatrix} \hat{p}_{A,MI^+}(\tau) \\ \hat{p}_{A,DE^+}(\tau) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \hat{p}_{A,MI^+DE^-}(\tau) \\ \hat{p}_{A,MI^-DE^+}(\tau) \\ \hat{p}_{A,MI^+DE^+}(\tau) \end{pmatrix} \tag{5.2.2}$$

In a similar manner the "competing risks" event type proportions can also be derived as

$$\begin{pmatrix} \hat{p}_{A,MI^+}(\tau) \\ \hat{p}_{A,MI^-DE^+}(\tau) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \hat{p}_{A,MI^+DE^-}(\tau) \\ \hat{p}_{A,MI^-DE^+}(\tau) \\ \hat{p}_{A,MI^+DE^+}(\tau) \end{pmatrix} \tag{5.2.3}$$

(and the same relations obviously also hold for group B).

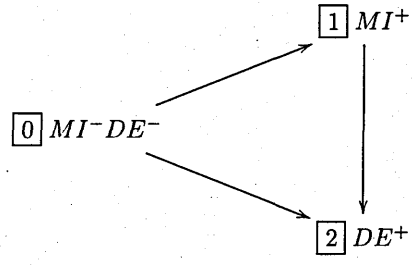
When right-censoring is present i.e. not all subjects are followed-up until time τ, the "exhaustive" event type proportions can no longer be estimated with simple proportions. Instead, estimates and corresponding asymptotic covariance matrices for $\hat{\mathcal{D}}(\tau)$ must be based on time-to-event modelling of the underlying multistate model. As discussed in books on multistate modelling (e.g. Andersen et al. (1997), chapter 2 and Beyersmann et al. (2012), chapter 8), the probabilities $\hat{p}_{A,k}(\tau)$ and $\hat{p}_{B,k}(\tau)$ in Formula (5.2.2) can then be nonparametrically estimated based on the Aalen-Johansen estimator for

the state transition probabilities of time-inhomogeneous Markov processes with a finite state space. Conveniently, the Aalen-Johansen estimator as well as a corresponding Greenwood type estimate of the associated covariance matrix have been implemented in the R package `etm` (Allignol et al. (2011)).

Using this method one can first specify a suitable multistate model for the setting of interest, and then easily derive the consistent estimates for all event type probabilities (and an associated covariance matrix) based on the corresponding transition probabilities. In what follows, I shall discuss how such a derivation is done for each setting of event types for the chosen illness-death model (which can be extended to more complex settings in a straightforward way).

Following the conventional multistate modelling notation, let $\hat{P}_{pq}(s,t)$ be the estimated transition probability of being at state q at time t given being at state p at time s with $s \leq t$. The multistate model suitable for the "marginal" setting in the current myocardial infarction and death illustration is depicted in Figure 5.2.1, where the relevant states are state 0: alive and MI-free, state 1: having an MI and state 2: death.

Figure 5.2.1: Multistate model for "marginal" setting.



Applying the aforementioned nonparametric multistate method to the multistate model in Figure 5.2.1, the "marginal" event type probabilities of interest in group A can be consistently estimated by the estimates of the transition probabilities between time 0 and time τ as follows:

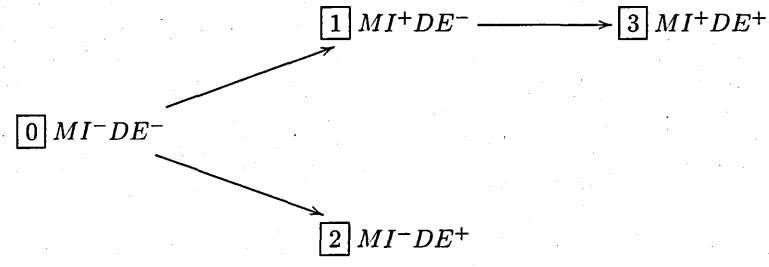
$$\hat{p}_{A,MI^+}(\tau) = \hat{P}_{01}(0,\tau), \text{ and } \hat{p}_{A,DE^+}(\tau) = \hat{P}_{02}(0,\tau)$$

The multistate model in Figure 5.2.1 could also be slightly modified for getting the "competing risks" event type probabilities by removing the transition from state 1 to state 2, and relabelling state 2 as MI^-DE^+. Of note, in this case the Aalen-Johansen estimator would reduce to the standard nonparametric CIF estimator discussed in Subsection 1.4.1.

Deriving the "exhaustive" event type probabilities using the Aalen-Johansen estimator for the model in Figure 5.2.1 is difficult as it is not clear how to disentangle the probabilities of MI^-DE^+ and MI^+DE^+. However, this computational issue can be resolved by explicitly accounting for the

states MI^-DE^+ and MI^+DE^+ via using the following multistate model

Figure 5.2.2: Multi-state model for "exhaustive" setting.



In this model, estimation of the "exhaustive" event type probabilities is as follows

$$\hat{p}_{A,MI^+DE^-}(\tau) = \hat{P}_{01}(0,\tau),\ \hat{p}_{A,MI^-DE^+}(\tau) = \hat{P}_{02}(0,\tau),\ \text{and}\ \hat{p}_{A,MI^+DE^+}(\tau) = \hat{P}_{03}(0,\tau)$$

Note that even though the "exhaustive" event type probabilities cannot be computed from the multistate model for "marginal" event types, the reverse is possible. This is because the linear relation in Formula (5.2.2) can also be used here to derive $(\hat{p}_{A,MI^+}, \hat{p}_{A,DE^+})$ from $\hat{p}_{A,MI^-DE^+}$, $\hat{p}_{A,MI^+DE^+}$ and $\hat{p}_{A,,MI^+DE^+}$.

5.3 Simultaneous inference for the weighted analysis of composite endpoints

Despite the appealing features of weighted composite endpoint analyses discussed so far, there exists a tangible difficulty in determining a single set of quantitative weights that is acceptable to all stakeholders. Sometimes instead of one, a finite set of weighting options are given (e.g. Hong et al. (2011) and Tong et al. (2012)). However, it is often even easier to rank or assign more qualitative inequality constraints to component outcomes or the event types derived from them. For example, rather than assigning quantitative weights to deaths (DE) and myocardial infarction (MI), it might be less controversial that the weight for a death event should be at least twice as large as the weight for a MI i.e. $w_{DE+} \geq 2w_{MI+}$, or even less controversially $w_{DE+} \geq w_{MI+}$. However, as an infinite number of weights fulfil such inequality constraints, multiple testing problems arise. Thus, a method which allows for the construction of confidence intervals for $\mathcal{T}(w,\tau)$ with (but not restricted to) simultaneous coverage of 95% across all weights and associated hypothesis test which protect the familywise type I error rate in the strong sense are desirable. In this section, I propose a multiplicity adjustment which achieves this.

This is an adaptation of method originally developed by Shapiro (2003) which is reviewed first

(Subsection 5.3.1) and then applied to weighted analyses of composite endpoints under the framework discussed in Section 5.2 (Subsection 5.3.2). As later shown, the proposed approach can incorporate a rich class of weight constraints including the nonnegativity constraints and an ordering constraints such as those mentioned above.

5.3.1 Simultaneous confidence intervals based on $\bar{\chi}^2$-distribution

By definition a set $\mathcal{C}$ in $\mathbb{R}^K$ is a cone if for all $w \in \mathcal{C}$ and for all $r > 0$, $rw \in \mathcal{C}$. For any chosen closed and convex cone $\mathcal{C}$ in $\mathbb{R}^K$, Shapiro (2003) suggested an approach for deriving simultaneous (two- or one-sided) confidence intervals with a desired overall coverage probability for the quantities $w^T\mathcal{D}$, where $w \in \mathcal{C}$ and $\mathcal{D} \in \mathbb{R}^K$ has a multivariate normal distribution $\mathcal{N}(\mathcal{D}_0, V)$ with V being a known nonsingular covariance matrix. In the following I first discuss one-sided confidence intervals with a simultaneous coverage of $1 - \frac{\alpha}{2}$ and then mention two-sided confidence intervals with an overall coverage of $1 - \alpha$.

For most practical purposes including the ones in this chapter, it is sufficient to consider cones $\mathcal{C}$ having the following form

$$\mathcal{C} = \left\{w \in \mathbb{R}^K : a_i^T w = 0, i = 1, \ldots, s; a_i^T w \geq 0, i = s+1, \ldots, K\right\} \tag{5.3.1}$$

where the $K \times K$ matrix $\left(a_1^T, \ldots, a_K^T\right)^T$ is of full rank and at least one of the constraints is an inequality constraint. For this setting, the one-sided confidence intervals of the following form are considered:

$$CI_L\left(w, V, c_{1-\frac{\alpha}{2}}\right) = \left[w^T\mathcal{D} - c_{1-\frac{\alpha}{2}}^{1/2}\left(w^T V w\right)^{1/2}, +\infty\right) \tag{5.3.2}$$

where $c_{1-\frac{\alpha}{2}}$ does not depend on each specific w but has to be chosen such that for given $\mathcal{C}$ and V, the corresponding simultaneous coverage probability is controlled at $1 - \frac{\alpha}{2}$ i.e.

$$P\left(\forall w \in \mathcal{C},\ w^T\mathcal{D}_0 \in CI_L\left(w, V, c_{1-\frac{\alpha}{2}}\right)\right) \geq 1 - \frac{\alpha}{2} \tag{5.3.3}$$

To achieve this, note that the event on the left-hand side is equivalent to

$$\max_{w \in \mathcal{C}} \frac{\left(\mathcal{D} - \mathcal{D}_0\right)^T w}{\left(w^T V w\right)^{1/2}} \leq c_{1-\alpha}^{1/2} \tag{5.3.4}$$

Consequently, the probability of simultaneous coverage becomes

$$P\left(\forall w \in \mathcal{C},\ w^T\mathcal{D}_0 \in CI_L\left(w, V, c_{1-\frac{\alpha}{2}}\right)\right) = P\left(\max_{w\in\mathcal{C}} \frac{(\mathcal{D}-\mathcal{D}_0)^T w}{(w^T V w)^{1/2}} \leq c_{1-\frac{\alpha}{2}}^{1/2}\right) \tag{5.3.5}$$
$$\geq P\left(\left[\max_{w\in\mathcal{C}} \frac{(\mathcal{D}-\mathcal{D}_0)^T w}{(w^T V w)^{1/2}}\right]^2 \leq c_{1-\frac{\alpha}{2}}\right)$$

For $Z(\mathcal{C}, V) = \max_{w\in\mathcal{C}} \frac{(\mathcal{D}-\mathcal{D}_0)^T w}{(w^T V w)^{1/2}}$, Shapiro (2003) showed that $[Z(\mathcal{C}, V)]^2$ has an exact chi-bar squared ($\bar{\chi}^2$) distribution. The theoretical properties of $\bar{\chi}^2$-distributions have been thoroughly discussed by many authors (e.g. Kudo (1963) and Shapiro (1988)). By definition, the distribution of a $\bar{\chi}^2$-variate is the weighted sum of independent χ^2-distributions. For $Z(\mathcal{C}, V)$ this means, for all c

$$P\left(Z(\mathcal{C}, V) \geq c\right) = \sum_{i=0}^{K} \tilde{w}_i\left(\mathcal{C}, V^{-1}\right) P\left(\chi_i^2 \geq c\right) \tag{5.3.6}$$

where χ_i^2 denotes a random variate following a central chi squared distribution with i degrees of freedom (with χ_0^2 defined as a point mass at 0 by convention), and $\tilde{w}_i\left(\mathcal{C}, V^{-1}\right)$, $i = 0, \ldots, K$ are the associated weights, summing to one, which can be derived from the given cone $\mathcal{C}$ and the inverse covariance matrix V^{-1}. It is a challenging task to analytically derive these quantities for any closed and convex cone $\mathcal{C}$. For detailed instructions on how to get $\tilde{w}_i\left(\mathcal{C}, V^{-1}\right)$, $i = 0, \ldots, K$ for the cones $\mathcal{C}$ defined in Equation (5.3.1) I refer to Section 5 of Shapiro (1988).

From the above, it is obvious that letting $c_{1-\frac{\alpha}{2}}$ be the $(1-\frac{\alpha}{2})$-quantile of the exact $\bar{\chi}^2$-distribution of $[Z(\mathcal{C}, V)]^2$ bounds the current overall coverage from below at $1-\frac{\alpha}{2}$.

To derive one-sided confidence intervals of the form

$$CI_U\left(w, V, c_{1-\frac{\alpha}{2}}\right) = \left(-\infty, w^T\mathcal{D} + c_{1-\frac{\alpha}{2}}^{1/2}\left(w^T V w\right)^{1/2}\right]$$

note that $\mathcal{D}-\mathcal{D}_0$ and $\mathcal{D}_0-\mathcal{D}$ have the same distribution and thus this is also true for the distributions of $\left[\max_{w\in\mathcal{C}} \frac{(\mathcal{D}_0-\mathcal{D})^T w}{(w^T V w)^{1/2}}\right]^2$ and $\left[\max_{w\in\mathcal{C}} \frac{(\mathcal{D}-\mathcal{D}_0)^T w}{(w^T V w)^{1/2}}\right]^2$. Hence, the same $\bar{\chi}^2$-distribution and the same critical values can be used as for CI_L.

Given the two one-sided simultaneously $1-\frac{\alpha}{2}$ confidence intervals, a simultaneous two-sided $1-\alpha$ confidence interval is given by

$$CI\left(w, V, c_{1-\frac{\alpha}{2}}\right) = \left[w^T\mathcal{D} - c_{1-\frac{\alpha}{2}}^{1/2}\left(w^T V w\right)^{1/2}, w^T\mathcal{D} + c_{1-\frac{\alpha}{2}}^{1/2}\left(w^T V w\right)^{1/2}\right]$$

5.3.2 Simultaneous inference in weighted composite endpoint analyses based on $\bar{\chi}^2$-distribution

Using Shapiro's method reviewed in the previous subsection for simultaneous inference under the framework discussed in Sections 5.1 - 5.2 is straightforward, except that the vector of differences in event type proportions $\hat{\mathcal{D}}(\tau)$ in Equation (5.2.1) only has a limiting normal distribution with a consistent covariance estimator $\hat{V}$. In fact, Shapiro (2003) did relax the exact normality assumption by allowing V to be known only up to a multiplicative factor $\sigma > 0$ i.e. $V = \sigma W$ if a consistent estimate for σ is available. However for the current purpose this is still a stringent restriction. In view of this, I provided in Appendix B.2 a proof that enables the asymptotic counterparts of the results in Subsection 5.3.1, of which the most crucial one is that if $\hat{\mathcal{D}}(\tau)$ follows an asymptotic normal distribution with an available consistent covariance estimator $\hat{V}$, then $\left[Z(\mathcal{C}, \hat{V})\right]^2 = \left[\max_{w \in \mathcal{C}} \frac{\left(\hat{\mathcal{D}}(\tau) - \mathcal{D}_0(\tau)\right)^T x}{\left(w^T \hat{V} w\right)^{1/2}}\right]^2$ has an asymptotic $\bar{\chi}^2$-distribution. Moreover, the finite-sample performance of this asymptotic approximation is investigated later on in Section 5.4.

Once these asymptotic results have been established, for a given $\hat{\mathcal{D}}(\tau)$ with the respective estimated (or exact) covariance matrix $\hat{V}$, simultaneous two- and one-sided confidence intervals with an overall coverage probability of $1 - \alpha$ and $1 - \frac{\alpha}{2}$, respectively, associated with $\mathcal{T}(w, \tau) = w^T \hat{\mathcal{D}}(\tau)$ for weight vectors $w \in \mathcal{C}$ (a closed and convex cone in $\mathbb{R}^K$) are

$$CI\left(w, \hat{V}, c_{1-\frac{\alpha}{2}}\right) = \left[w^T \hat{\mathcal{D}}(\tau) - c_{1-\frac{\alpha}{2}}^{1/2} \left(w^T \hat{V} w\right)^{1/2}, w^T \hat{\mathcal{D}}(\tau) + c_{1-\frac{\alpha}{2}}^{1/2} \left(w^T \hat{V} w\right)^{1/2}\right] \text{ and} \quad (5.3.7)$$

$$CI_L\left(w, \hat{V}, c_{1-\frac{\alpha}{2}}\right) = \left[w^T \mathcal{D} - c_{1-\frac{\alpha}{2}}^{1/2} \left(w^T V w\right)^{1/2}, +\infty\right), \text{ or} \quad (5.3.8)$$

$$CI_U\left(w, \hat{V}, c_{1-\frac{\alpha}{2}}\right) = \left(-\infty, w^T \mathcal{D} + c_{1-\frac{\alpha}{2}}^{1/2} \left(w^T V w\right)^{1/2}\right], \text{ respectively}$$

Using the duality between hypothesis testing and confidence intervals, these simultaneous confidence intervals can also be used to derive associated tests which control the familywise type I error rate in the strong sense across the two-sided or one-sided null hypotheses $H_0 : w^T \hat{\mathcal{D}}(\tau) = 0$ or $H_0 : w^T \hat{\mathcal{D}}(\tau) \leq 0$, respectively, with weights fulfilling the desired constraints. Specifically, for any specific set of weights fulfilling the desired constraints, the associated hypothesis test can be tested by checking whether 0 is contained in the associated simultaneous confidence interval. This leads to the following practical advantage. Those who agree on restricting the weight vector w to $\mathcal{C}$ may well still have varying preferences for an exact w. Each of them can then test and possibly reject the null hypothesis induced by their weight vectors of choice without worrying about inflating the overall type I error.

Finally, as previously mentioned in Subsection 5.2.1, forcing the weight vectors to have elements summing to one can assist with the interpretation of $w^T \hat{\mathcal{D}}(\tau)$. Strictly speaking there exists no cone

satisfying such a constraint. However, this constraint can simply be ignored in all of the above calculations because any weight vector w can be normalised to $\frac{w}{\sum_{i=1}^{K} w_i}$ which only scales the corresponding weighted statistic and its confidence interval by the same amount, and thus does not affect coverage probability.

5.4 Simulation studies

This section demonstrates the finite-sample performance of applying the simultaneous inference strategy proposed in Section 5.3 via a series of simulation scenarios. The simulation scenarios mimic an RCT with outcomes following an illness-death model, similar to what has been used for illustration throughout this chapter. In analogy to the previous sections, I will also refer to the transient state as MI and the absorbing state as DE. More details about the simulation scenarios are given in the following subsection.

5.4.1 Scenarios

Each scenario has two groups of patients representing two treatment arms in an RCT whose maximum follow-up duration is set at time $\tau = 5$ (years). Data in each treatment group were generated from the multistate model according to the "exhaustive" setting with the following states: state 0-MI^-DE^-, state 1-MI^+DE^-, state 2-MI^-DE^+ and state 3-MI^+DE^+. For this simulation I used constant transition hazards λ_{rs}, $rs \in \{01, 02, 13\}$ whose chosen values are displayed in Table 5.1 for various treatment groups. The treatment arms are configured as follows: A represents a common control treatment for all interventions, B is a new intervention that has no effect i.e. has exactly the same transition hazards as A, C is a treatment reducing every transition hazard by 25%, and finally D is a treatment that only reduces the transition hazard λ_{01} by 25%. In total I examined 6 pairs of treatment arms corresponding to two different sets of transition hazards for the control treatment A and the intervention effects as described above.

Besides employing scenarios with only administrative right-censoring at time $\tau = 5$ as above, I also simulated additional scenarios which include both administrative right-censoring and independent right-censoring following an exponential distribution with rate $\lambda = 0.05$. The resulting censoring probabilities caused by this extra censoring mechanism corresponding to each treatment arm are given in the bottom half of Table 5.1. In addition, this table also displays the "exhaustive" and the resulting "marginal" event type probabilities under each treatment arm and censoring scheme.

Finally I varied the sample size between 100 and 500 per group which together with the above factors led to 24 scenarios. Results for each of the 24 scenarios are based on 1000 simulated data sets.

Figure 5.4.1: Multistate model for "exhaustive" setting.

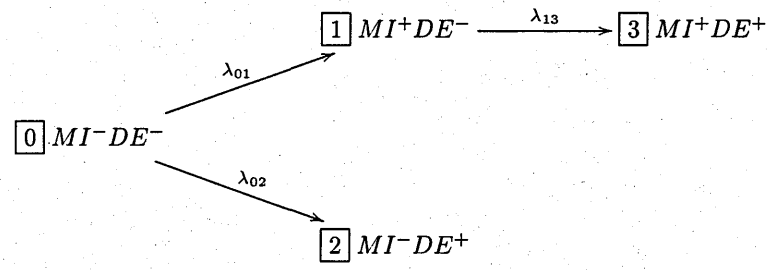


Table 5.1: Transition hazards (λ_{rs}) w.r.t. Figure 5.4.1 and resulting event type probabilities with and without right-censoring.

	No right-censoring before τ									
	$\lambda_{rs} \times 10^{-2}$			$p_{ek}(\tau)^*$ %				$p_{mk}(\tau)^\$$ %	
rs / Group	01	02	13	k	MI^+DE^-	MI^-DE^+	MI^+DE^+	k	MI^+	DE^+
A_1 & B_1	5.00	5.00	30.00		9.6	19.7	10.1		19.7	29.8
C_1	3.75	3.75	22.50		9.1	15.7	6.6		15.7	22.3
D_1	3.75	5.00	30.00		7.4	20.3	7.7		15.1	28.0
A_2 & B_2	5.00	2.00	20.00		12.9	8.4	8.2		21.1	16.6
C_2	3.75	1.50	15.00		11.4	6.6	5.1		16.5	11.7
D_2	3.75	2.00	20.00		10.1	8.7	6.2		16.3	14.9
	Right-censoring before $\tau^@$									
A_1 & B_1		19.0$^{\$\$}$			8.9	17.6	8.6		17.5	26.2
C_1		19.7$^{\$\$}$			8.3	13.9	5.6		13.9	19.5
D_1		19.2$^{\$\$}$			6.9	18.1	6.6		13.5	24.7
A_2 & B_2		20.6$^{\$\$}$			11.8	7.5	7.0		18.8	14.5
C_2		21.1$^{\$\$}$			10.4	5.8	4.3		14.3	10.1
D_2		20.6$^{\$\$}$			9.2	7.8	5.4		14.6	13.2

*: "Exhaustive" setting, $: "Marginal" setting.
$$: Overall censoring probability (%) before $\tau = 5$.
@: the associated event probabilities display the average observed frequency of those event types, i.e. subjects censored before τ were counted as not having an event for the purpose of this table.

5.4.2 Weight constraints and competing approaches

In all simulation scenarios I considered nonnegativity constraints i.e. all elements in a weight vector must be nonnegative, as well as the following ordering constraint for the current "exhaustive" settings

$$0 \leq w_{MI^+DE^-} \leq w_{MI^-DE^+} \leq w_{MI^+DE^+}$$

Since the "marginal" event type probabilities are readily available for each "exhaustive" setting, I also applied the same weight constraints to these probabilities for which the ordering constraint is

$$0 \leq w_{MI+} \leq w_{DE+}$$

For the setting of administrative censoring only, confidence intervals were based on observed proportions experiencing each event type and the asymptotic covariance matrix was based on the multinomial distribution. For settings with censoring the Aalen-Johansen estimator and an associated Greenwood-type estimate of the covariance matrix was used as described in Subsection 5.2.3.

In addition to the method for constructing simultaneous confidence intervals based on the $\bar{\chi}^2$-distribution (hereinafter abbreviated by $\bar{\chi}^2$-method) proposed in Subsection 5.3.2, I considered the following competing approaches: an asymptotic version of Scheffe's method (Scheffe (1953)) (which guarantees simultaneous control across all weight vectors without any constraints) and unadjusted confidence intervals without any multiplicity adjustment.

5.4.3 Assessment methods

For each simulation scenario, I studied the performance of the competing methods for both the "exhaustive" and the associated "marginal" settings under the weight constraints mentioned above. Specifically the considered two-sided 95%-confidence intervals which, for the $\bar{\chi}^2$-method, Scheffe's method and the unadjusted method, respectively, take the following forms:

$$CI\left(w, \hat{V}, c_{1-\frac{\alpha}{2}}\right) = \left[w^T\hat{\mathcal{D}}(\tau) - c_{1-\frac{\alpha}{2}}^{1/2}\left(w^T\hat{V}w\right)^{1/2}, w^T\hat{\mathcal{D}}(\tau) + c_{1-\frac{\alpha}{2}}^{1/2}\left(w^T\hat{V}w\right)^{1/2}\right] \text{ and}$$

$$CI\left(w, \hat{V}, s_{1-\alpha}\right) = \left[w^T\hat{\mathcal{D}}(\tau) - s_{1-\alpha}^{1/2}\left(w^T\hat{V}w\right)^{1/2}, w^T\hat{\mathcal{D}}(\tau) + s_{1-\alpha}^{1/2}\left(w^T\hat{V}w\right)^{1/2}\right] \text{ and}$$

$$CI\left(w, \hat{V}, q_{1-\frac{\alpha}{2}}\right) = \left[w^T\hat{\mathcal{D}}(\tau) - q_{1-\frac{\alpha}{2}}\left(w^T\hat{V}w\right)^{1/2}, w^T\hat{\mathcal{D}}(\tau) + q_{1-\frac{\alpha}{2}}\left(w^T\hat{V}w\right)^{1/2}\right]$$

Following the terminology of Subsections 5.3.2 and 5.2.1 $c_{1-\frac{\alpha}{2}}$ is the 97.5% quantile of the $\bar{\chi}^2$-distribution corresponding to the approximate covariance matrix and cone constraint, $s_{1-\alpha}$ is the 95% quantile of the χ^2-distribution (with 2 and 3 degrees of freedom for the "marginal" and "exhaustive" settings, respectively) corresponding to an asymptotic Scheffe's method, and $q_{1-\frac{\alpha}{2}}$ (≈ 1.96) is the 97.5% quantile of the standard normal distribution required for the unadjusted method.

As the main goal of the proposed method is to maintain a nominal simultaneous coverage probability of 95%, the first evaluation criterion is the Monte Carlo simultaneous coverage probability across all weight vectors in the cone defined by each constraint. However, there are an infinite number of

points inside a cone. To approximate the true simultaneous coverage, I evaluated coverage across a grid of uniformly spaced points across the intersection of the cone with the hyperplane defined by the weight constraint $\sum_{i=1}^{K} w_i = 1$. Specifically, I chose 2463 and 500 grid points for the "exhaustive" and "marginal" settings, respectively, under the nonnegativity contstraint. For the ordering constraint, I included the subset of the above grid points satisfying that additional constraint.

Second, a commonly known trade-off for controlling the simultaneous coverage is the loss of efficiency i.e. the resulting confidence intervals are more conservative (wider) than the associated unadjusted one. In view of this, I compared the $\bar{\chi}^2$-method and Scheffe's method to the unadjusted one via two Monte Carlo relative efficiency type measures defined, respectively, as $\frac{c_{1-\frac{\alpha}{2}}^{1/2}}{q_{1-\frac{\alpha}{2}}}$ and $\frac{s_{1-\alpha}^{1/2}}{q_{1-\frac{\alpha}{2}}}$. It is easy to see that these quantities are the ratios of the relevant confidence interval widths.

5.4.4 Results

Simulation results for the "exhaustive" and "marginal" settings in each scenario are displayed in Tables 5.2 and 5.3, respectively. Observed coverage for the $\bar{\chi}^2$-method was very close to the nominal 95%. Coverage was larger than 93% for all settings including those involving right-censoring. Coverage more than two Monte-Carlo standard errors below 95%, i.e. coverage below 93.6%, occurred for 3/96 reported coverage probabilities, all of which involved intervention C_1. As expected, Scheffe's method yielded coverage beyond the nominal level whereas the unadjusted confidence intervals had simultaneous undercoverage.

Monte Carlo relative efficiency (first three result columns in Table 5.2) followed the anticipated pattern that relative to the unadjusted method the $\bar{\chi}^2$-method is more efficient than Scheffe's method, especially under the more restrictive constraint as Scheffe's method disregards these weight restrictions. Compared to the unadjusted confidence intervals, confidence intervals for the $\bar{\chi}^2$-method under an ordering constraint were approximately 22% wider for the exhaustive and 10% wider for the marginal setting. In contrast, Scheffe's method led to increases by 43% and 25%.

Table 5.2: Simulation results for "exhaustive" settings.

Groups	n	No right-censoring before τ								
		MC Relative efficiency			MC Simultaneous coverage %[@]					
		$\bar{\chi}^{2*}$		Scheffe[$]	$\bar{\chi}^{2*}$		Scheffe[$]		Unadjusted[#]	
Weight constraints		$\mathbb{R}^3_+$	$\leq$		$\mathbb{R}^3_+$	$\leq$	$\mathbb{R}^3_+$	$\leq$	$\mathbb{R}^3_+$	$\leq$
A_1&B_1	100	1.37	1.21	1.43	94.5	95.1	96.0	98.1	77.3	86.7
	500	1.37	1.21	1.43	94.0	95.7	95.8	98.3	75.8	87.1
A_1&C_1	100	1.36	1.22	1.43	93.7	94.2	95.0	97.5	73.8	84.6
	500	1.36	1.22	1.43	93.5	93.3	96.0	98.1	75.1	83.7
A_1&D_1	100	1.36	1.21	1.43	95.1	94.9	96.5	98.5	76.6	87.2
	500	1.36	1.21	1.43	94.6	95.0	96.0	97.9	76.7	86.4
A_2&B_2	100	1.36	1.21	1.43	95.4	95.5	96.6	98.4	76.7	88.1
	500	1.36	1.21	1.43	94.8	96.2	96.3	98.5	73.4	85.7
A_2&C_2	100	1.35	1.22	1.43	93.7	93.8	95.3	97.5	76.0	86.7
	500	1.36	1.22	1.43	94.4	94.0	95.6	98.3	76.8	87.5
A_1&D_2	100	1.36	1.21	1.43	93.9	93.6	95.6	98.0	76.2	85.7
	500	1.36	1.21	1.43	94.9	95.0	96.3	98.3	78.9	88.3
		Right-censoring before τ								
A_1&B_1	100	1.37	1.22	1.43	93.7	94.3	95.0	98.0	74.0	84.8
	500	1.37	1.22	1.43	96.4	95.7	97.4	98.8	79.8	88.5
A_1&C_1	100	1.36	1.22	1.43	95.1	95.4	96.3	97.8	77.8	87.8
	500	1.36	1.22	1.43	94.1	94.2	95.7	97.7	76.6	86.6
A_1&D_1	100	1.36	1.22	1.43	94.5	95.1	95.5	98.3	75.1	87.2
	500	1.36	1.22	1.43	95.4	95.3	96.6	98.5	75.5	87.0
A_2&B_2	100	1.36	1.21	1.43	94.8	94.0	96.5	97.8	78.2	87.5
	500	1.36	1.22	1.43	94.9	95.6	96.2	98.2	78.3	86.7
A_2&C_2	100	1.36	1.22	1.43	95.2	95.4	96.6	98.3	76.8	86.4
	500	1.36	1.22	1.43	95.6	95.5	96.6	98.0	77.1	87.0
A_1&D_2	100	1.36	1.22	1.43	95.2	94.9	96.8	98.7	76.9	88.3
	500	1.36	1.22	1.43	95.5	95.6	96.7	98.1	75.9	86.9

*: resulting from simultaneous confidence interval based on $\bar{\chi}^2$ method.
$: resulting from simultaneous confidence interval based on Scheffe's method.
#: resulting from unadjusted simultaneous confidence interval.
$\mathbb{R}^3_+$: nonnegativity constraint, $\leq$: ordering constraint.
@: Monte Carlo standard error $\approx$ 0.7%.

Table 5.3: Simulation results for "marginal" settings.

Groups	n	No right-censoring before τ								
		MC Relative efficiency			MC Simultaneous coverage %[@]					
		$\bar{\chi}^{2*}$		Scheffe[$]	$\bar{\chi}^{2*}$		Scheffe[$]		Unadjusted[#]	
Weight constraints		$\mathbb{R}^2_+$	$\leq$		$\mathbb{R}^2_+$	$\leq$	$\mathbb{R}^2_+$	$\leq$	$\mathbb{R}^2_+$	$\leq$
A_1&B_1	100	1.18	1.10	1.25	94.3	94.2	95.7	97.3	89.0	91.7
	500	1.18	1.10	1.25	95.6	96.0	96.6	97.9	89.4	93.0
A_1&C_1	100	1.18	1.10	1.25	94.5	94.5	96.1	97.3	86.7	90.6
	500	1.18	1.10	1.25	93.9	93.4	95.1	96.9	86.7	90.6
A_1&D_1	100	1.18	1.10	1.25	95.2	95.0	96.6	97.8	89.6	92.5
	500	1.18	1.10	1.25	94.8	94.8	95.9	97.1	88.4	91.6
A_2&B_2	100	1.17	1.11	1.25	95.8	95.7	96.6	97.8	89.9	93.1
	500	1.17	1.11	1.25	95.6	95.8	97.9	98.7	88.0	91.4
A_2&C_2	100	1.17	1.11	1.25	93.2	93.7	94.8	95.9	88.1	90.7
	500	1.17	1.11	1.25	93.9	94.5	96.1	97.3	89.4	91.8
A_1&D_2	100	1.17	1.11	1.25	93.9	93.6	95.7	96.7	87.3	90.0
	500	1.17	1.11	1.25	95.3	95.2	97.1	97.8	88.9	91.5
		Right-censoring before τ								
A_1&B_1	100	1.18	1.10	1.25	93.7	95.2	95.7	97.9	86.5	91.1
	500	1.18	1.10	1.25	95.7	95.4	97.5	98.3	90.3	92.5
A_1&C_1	100	1.18	1.10	1.25	95.8	95.1	96.5	97.4	89.3	92.7
	500	1.18	1.10	1.25	94.2	94.5	96.2	97.9	87.9	91.5
A_1&D_1	100	1.18	1.10	1.25	94.1	94.6	96.2	97.8	88.5	92.8
	500	1.18	1.10	1.25	95.2	95.5	96.6	97.4	89.8	92.4
A_2&B_2	100	1.17	1.11	1.25	94.1	94.2	96.2	97.3	88.7	91.1
	500	1.17	1.11	1.25	95.5	95.5	96.6	97.5	89.7	92.8
A_2&C_2	100	1.18	1.11	1.25	95.9	95.2	97.2	97.8	88.6	91.7
	500	1.18	1.11	1.25	95.6	95.3	96.9	97.9	88.8	91.8
A_1&D_2	100	1.17	1.11	1.25	95.4	95.4	96.6	97.1	90.1	92.7
	500	1.18	1.11	1.25	95.3	95.4	96.5	97.4	89.3	92.9

*: resulting from simultaneous confidence interval based on $\bar{\chi}^2$ method.
$: resulting from simultaneous confidence interval based on Scheffe's method.
#: resulting from unadjusted simultaneous confidence interval.
$\mathbb{R}^2_+$: nonnegativity constraint, $\leq$: ordering constraint.
@: Monte Carlo standard error $\approx 0.7\%$.

5.5 Applications

5.5.1 Design consideration for a cardiovascular trial

In this case study, I apply the proposed methods to the design of a hypothetical RCT in cardiology with a composite endpoint. This illustration is based on the work of Hong et al. (2011), who suggested to assign weights corresponding to standardized disability-adjusted life-years (DALY) to the following three common outcomes in vascular prevention trials: nonfatal stroke (ST), nonfatal myocardial

infarction (MI) and vascular death (DE). These authors only considered first events i.e. the "competing risks" setting (see Subsection 5.2.2), and reported the weights for three age categories which are reproduced in Table 5.4 below for convenience.

Table 5.4: DALY lost for first vascular events (according to Table 2 of Hong et al. (2011)).

	age 50	age 60	age 70
Nonfatal Stroke (ST^+_{1st})	10.49	7.63	5.06
Nonfatal MI (MI^+_{1st})	6.73	5.14	3.85
Vascular Death (DE^+_{1st})	16.79	11.59	7.24

For ease of later discussion let each column in Table 5.4 form the respective weight vectors w_{50}, w_{60} and w_{70}. I further assume that the RCT is a 1:1 randomized trial and the assumed 3-year risks for the 3 event types are provided in Table 5.5. Based on this information the aim is to calculate the sample size of the trial under the following varying sets of requirements:

1. 90% power to detect an effect on the composite endpoint (any of the first events) at the unadjusted (two-sided) 5% significance level. This corresponds to a conventional analysis and sample size calculation which assumes an absolute risk reduction of the composite endpoint from 30% to 20%.

2. 90% power to detect an effect on the composite endpoint at the multiplicity adjusted 5% significance level. Multiplicity adjustment to guarantee strong control of the family-wise type I error across all possible linear weight combinations, i.e. $\mathcal{C}_2 = \mathbb{R}^3$.

3. 90% power to detect an effect on the composite endpoint at the multiplicity adjusted 5% significance level. Multiplicity adjustment to guarantee strong control of the family-wise type I error across all linear weight combinations with non-negative weights, i.e. $\mathcal{C}_3 = \mathbb{R}^3_+$.

4. 90% power to detect an effect on the composite endpoint at the multiplicity adjusted 5% significance level. Multiplicity adjustment to guarantee strong control of the family-wise type I error across all non-negative weights following an ordering constraint, i.e.

$$\mathcal{C}_4 = \left\{ w \in \mathbb{R}^3 : \; w_{DE^+_{1st}} \geq w_{ST^+_{1st}} \geq w_{MI^+_{1st}} \right\}$$

5. A simultaneous power of 90% to detect an intervention effect for all weighted differences corresponding to weights in the cone spanned by the DALY weight vectors (columns) given in Table 5.4, i.e.

$$\mathcal{C}_5 = \left\{ w \in \mathbb{R}^3 : \; w = \alpha_1 w_{50} + \alpha_2 w_{60} + \alpha_3 w_{70}, \; \forall (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{R}^3_+ \right\}$$

and control of the familywise type I error at 5% across this cone.

Table 5.5: Assumed 3-year risks.

Requirement	Control	Intervention
Nonfatal Stroke (ST^+_{1st})	15%	10%
Nonfatal MI (MI^+_{1st})	10%	7.5%
Vascular Death (DE^+_{1st})	5%	2.5%

The resulting sample sizes are reported in Table 5.6. For the first 4 requirements, sample size was calculated based on standard sample size formulas for 2-group comparisons. The corresponding critical values for settings 1 and 2 are 1.96 (the 97.5% quantile of the normal distribution) and 2.79 (i.e. the square-root of the 95% quantile of a chi-squared distribution with 3 degrees of freedom required for an asymptotic Scheffe-correction), respectively. For settings 3 and 4 the respective critical values were based on the relevant $\bar{\chi}^2$-distributions. For requirement 5 the sample size calculation was based on repeated simulations with increasing sample sizes until the simultaneous power was achieved across a set of 918 normalised weight vectors w equally spaced across the cone under consideration as previously described in 5.4.3.

Table 5.6: Sample size result.

Requirement	Sample size (per group)	Relative efficiency*
1	392	1
2	621	1.43
3	519	1.25
4	488	1.19
5	381	1.02

Note: Bonferroni-correction for composite and component endpoints gives a critical value $\approx 1.27 \times 1.96$. *: compared to 1.96.

Table 5.6 exhibits the foreseen pattern that a larger sample size is required if multiplicity adjustment is required across larger sets of weights. This also shows how inefficient Scheffe's method can be in cases of more restrictive constraints. For example, to fulfil the ordering constraint of C_4 the $\bar{\chi}^2$-method only increases the sample size by about 25% while Scheffe's method would inflate the sample size by almost 60%. Of note, a simple Bonferroni correction which adjusts for multiplicity for only 4 comparisons (e.g. an analysis of the composite endpoint and the 3 component outcomes), would lead to a critical value of 1.27×1.96. Accounting for such a Bonferroni adjustment would lead to a larger inflation in sample size than requirements 4 while formally only guaranteeing multiplicity control across a much smaller set of null hypotheses.

Interestingly, for the fifth requirement the $\bar{\chi}^2$- method requires a smaller sample size than the standard approach. This seemingly peculiar result can be explained by the following two factors.

First the weight vectors w_{50}, w_{60} and w_{70} all give lowest weight to nonfatal MI whose assumed intervention effect was smallest. Second, despite being quantitatively different these weight vectors span a "narrow" cone, hence requiring only minimal multiplicity adjustment.

5.5.2 Weighted analysis of a composite endpoint in a trial of uncomplicated enteric fever

This example illustrates the proposed method to data from a typhoid trial conducted at the Nepal site of OUCRU-VN that studied the efficacy of two antibiotic treatments for uncomplicated enteric fever: Gatifloxacin and Cefiximie (Pandit et al. (2007)). The outcome of interest here is the composite endpoint of overall treatment failure which was a secondary endpoint in the trial. Overall treatment failure was defined as acute treatment failure (severe complications, fever or other persistent symptoms for more than 7 days, or requirement for rescue treatment), death or relapse (fever with a positive blood culture within a month of completing treatment). Only one death occurred and it was pooled with acute treatment failure for the sake of this example. By definition, subjects with an acute treatment failure were not evaluated for relapse, hence the event types are exclusive. Overall results for patients with culture-confirmed typhoid are displayed in Table 5.7.

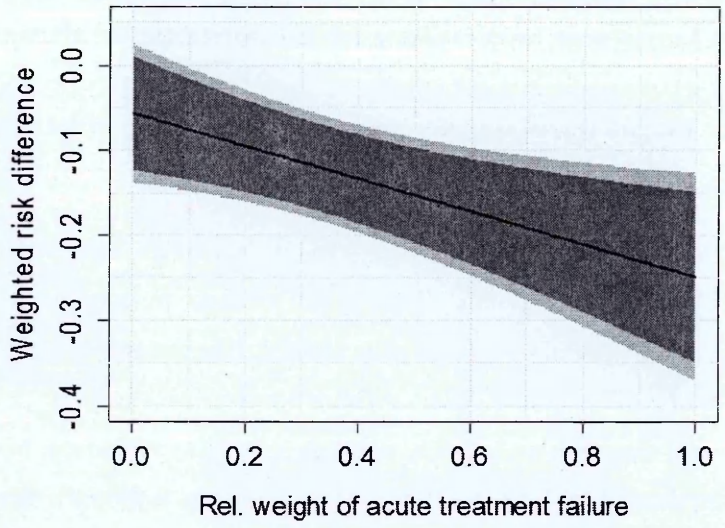
Table 5.7: Frequencies of component outcomes in patients with culture-confirmed typhoid (based on Figure 1 in Pandit et al. (2007)).

	Gatifloxacine (n=92)	Cefixime (n=77)
Acute treatment failure or death	1 (1.09%)	20 (25.97%)*
Relapse	2 (2.17%)	6 (7.79%)

*: including one death.

From Table 5.7 it is clear that Gatifloxacine strongly reduced the risk of experiencing an acute treatment failure or death compared to Cefixime whereas the effect on relapse is much less pronounced. As it is unclear which of the two event types is more clinically relevant in general, I did not impose an ordering constraint but instead adjusted for multiplicity taking into account only the nonnegativity constraint. Figure 5.5.1 shows the weighted test statistic $T(w, \tau)$ and associated 95% confidence intervals (with and without multiplicity adjustment) depending on the weight assigned to "acute treatment failure or death" and assuming that the two weights are standardized to sum up to 1.

Figure 5.5.1: Weighted risk difference depending on the relative weight of "acute treatment failure or death".



Black line: Weighted risk difference. Dark gray: Unadjusted 95% CI. Light gray: Simultaneous 95% CI.

According to Figure 5.5.1, across all weight choices the associated 95%-confidence intervals given by the $\bar{\chi}^2$- method are marginally wider than the unadjusted ones. Moving from 0 to 1 on the relative weight horizon, the simultaneous 95%-confidence intervals based on the $\bar{\chi}^2$- method begins to claim a significant difference in the weighted risks as soon as the relative weight of acute treatment failure or death is greater than 10%. As relative weights below 10% for this component outcome seem unreasonable, the trial demonstrates superiority of Gatifloxacin over Cefixime with respect to the combined outcome simultaneously across all "clinically reasonable" weights. Moreover, in this illustration the cost for multiplicity adjustment using the $\bar{\chi}^2$- method is only modest.

5.6 Discussions

In this chapter I developed a new approach to the weighted analysis of composite endpoints that can handle both binary and time-to-event data. The main idea is to consider the weighted risk differences of the component event types of interest where weights can be assigned to component event types in flexible ways. I further proposed to use Shapiro's $\bar{\chi}^2$-method (Shapiro (2003)) to construct simultaneous confidence intervals for the true weighted differences and associated tests that protect the familywise type I error in the strong sense across all weights inside a chosen closed and convex cone. To my knowledge, this chapter is the first substantive application of Shapiro's method to a practical problem where constraints naturally arise and the first evaluation of the method in a simulation study. I also extended the results of Shapiro (2003) to the case where the relevant random variables follow

normal distributions only asymptotically which was not mentioned in his original work.

Basing the proposed weighted analysis framework on Shapiro's $\bar{\chi}^2$-method allows to handle a flexible class of weight constraints that can be expressed as a system of linear equality and inequality constraints. Such a class is rich enough for most practical purposes and thus addresses the problems in choosing exact weights as well as the general interpretation issue involving composite endpoints. Moreover, the efficiency loss in exchange for multiplicity adjustment induced by the proposed method is modest compared to the method based on Scheffe's procedure (Scheffe (1953)) as demonstrated in both simulations and applications. Of note, Shapiro (2003) and Shapiro (1988) also discussed more general cone constraints. However, it is unclear how to derive the exact $\bar{\chi}^2$-distributions in most other cases.

For composite events, it is usually sensible to assign nonnegative weights to all component outcomes as all of them are harmful. However there is no technical barrier that prevents using the developed framework for weighted analyses that trade off more general beneficial and harmful outcomes by employing weights of opposite signs. In addition, even though only effects on the absolute risk scale which is most relevant to clinical decision making were discussed, the proposed approach could also be applied to the joint analysis of any set of multiple endpoints (with treatment effects expressed e.g. on the log-hazard ratio scale) as long they follow a joint asymptotic normal distribution.

A possible alternative approach to the analysis of composite time-to-event endpoints would be to use an integrated version of the test statistic proposed in Equation (5.2.1) which could be defined as:

$$\mathcal{T}^*(w,\tau) = \sum_{k=1}^{K} w_k \int_0^{\tau} h(t)\left(\hat{p}_{A,k}(t) - \hat{p}_{B,k}(t)\right) dt$$

where $h(\cdot)$ is a deterministic function of time which can be used to emphasize on the importance of absolute risk difference over different time periods. I did not pursue this further because in clinical applications, justifying the exact choice of $h(\cdot)$ would often be difficult and obtaining a covariance matrix for $\mathcal{T}(w,\tau)$ is not straightforward with standard software.

Another possible extension of the proposed approach is to consider test statistics that adjust for covariates other than treatment assignment, where the simplest starting point is a stratified variant of the test statistic $\mathcal{T}(w,\tau)$.

Finally the current work focused on Wald-type confidence intervals which are known to be unreliable if event probabilities and sample size are low even in simple settings (DasGupta et al. (2001)). Thus extensions of the current approach to likelihood-ratio- or score-based confidence intervals would be desirable and could be a potential area for future research.

Chapter 6

Overview and outlook

This thesis made several methodological contributions to the analysis of competing risks data and composite endpoints. This final chapter briefly summarizes the contribution of my work to these fields and discusses some potential areas for future research.

6.1 Contributions

6.1.1 A flexible model for the estimation of cumulative incidence functions

The cumulative incidence function (CIF) describes how the absolute risk of experiencing a specific event type changes over time in the presence of other competing events and is one of the most important quantities in competing risks.

In Chapter 3 I proposed a novel semi-nonparametric (SNP) method for CIF-estimation based on earlier work on SNP density estimation (Gallant & Nychka (1987) and Zhang & Davidian (2008)). I presented the relevant likelihood calculations, developed a greedy stepwise forward algorithm for estimation and model selection and implemented it in the statistical software R. The proposed method combines the strength of existing parametric and nonparametric approaches in the sense that it is applicable under arbitrary censoring and truncation without imposing stringent parametric restrictions.

A rigorous justification for the asymptotic properties of the proposed model was not possible. However, I conducted an extensive simulation study which demonstrated that the proposed method compares favourably to competing parametric and nonparametric approaches in terms of accuracy and that confidence intervals based on ad-hoc asymptotic inference have the expected finite-sample coverage in many situations.

6.1.2 A flexible model for CIF-based regression

In Chapter 4, I extended the model for CIF estimation to regression modelling aiming to estimate covariate-dependent CIF estimates. To my knowledge this is the first flexible regression model for competing risks that can handle interval-censored data. The accompanying simulation studies showed that this regression model is competitive compared to alternative approaches in terms of prediction accuracy but that for the validity of ad-hoc asymptotic inference a relatively large sample size is required. The model is based on a mixture factorization and shares limitations related to identifiability and interpretation with similar models which are discussed in detail. To assess the adequacy of a fitted model, I also proposed a method for regression diagnostics applicable to any competing risks models based on the mixture factorization.

6.1.3 Weighted analyses of composite endpoints

Composite endpoints are widely used in randomized controlled trials but the classical analysis of composite endpoints has several disadvantages including only considering the first occurring event and weighting all component outcomes equally despite the fact that they may differ in their clinical importance. In Chapter 5, I present a unified framework for the weighted comparison of both binary and time-to-event composite endpoints between treatment groups. The proposed test statistic can be interpreted as a weighted average risk difference or, if weights correspond to costs, as an expected difference in costs. Exact weights are often difficult to obtain and I present a method for multiplicity control across sets of weight vectors which satisfy a flexible set of inequality and equality constraints. This generalized work on the $\bar{\chi}2$-distribution by Shapiro (Shapiro (2003)) to the asymptotic setting, where test statistics only approximately follow a normal distribution, and applies it to a setting where such constraints naturally arise.

6.2 Outlook

6.2.1 Asymptotic properties of SNP methods

As described in Chapters 3 and 4, proofs of consistency and asymptotic normality of the proposed SNP estimators of the CIF and CIF-based regression models are still lacking. Of note, previous work on the usage of SNP densities in survival models also did not establish mathematical properties of the proposed estimators (Zhang & Davidian (2008)). If successfully established, an asymptotic theory for these SNP methods would provide a solid justification for the construction of confidence intervals and statistical tests. However, this is a challenging task as consistency and asymptotic normality for

sieve maximum likelihood estimators, of which our model is a special case, are in general difficult to verify even in situations that are simpler than those considered in competing risks and survival analysis (Bierens (2014)).

Consistency has been established for the simpler case of SNP estimation of a density function (Fenton & Gallant (1996*b*)) based on the conditions given by Gallant & Nychka (1987). Accordingly, a possible strategy to attack the consistency problem is to extend the consistency results of Gallant & Nychka (1987) first to survival analysis (where censoring and the introduction of covariates provide challenges) and subsequently to competing risks. For the proposed competing risks model, this is further complicated by the fact that it is based on a mixture factorization which poses identifiability problems when the observed follow-up duration is limited.

Furthermore, to my knowledge, no publication has yet formally demonstrated the asymptotic normality of estimators based on SNP densities. Hence such a proof would be even more challenging.

6.2.2 Faster estimation algorithm for SNP methods

The proposed algorithm for CIF estimation and CIF-based regression appeared to work reliably in the simulation studies in Chapters 3 and 4. However, despite utilising parallel computing empowered by package `parallel` for the statistical software R, the algorithm requires substantial computing power. A faster algorithm would be highly desirable for several reasons: It would allow for more extensive simulation studies which could provide better understanding about the asymptotic performance of the SNP method and make the method more attractive for practical usage. It would also allow the fitting of models with higher SNP polynomial degrees which, as illustrated in Subsection 3.6.3, might lead to more accurate fits. Finally, it would allow for routine use of bootstrap-based inference which could potentially be more reliable.

6.2.3 Alternative competing risks models based on SNP densities

One limitation of the proposed competing risk regression model is that it relies on the mixture factorization and thus on the event status at time infinity which may be poorly identified based on the available data. In Section 4.4, I briefly described a potential alternative model which conditions on the event status at a finite follow-up time point instead. Developing this approach would require the flexible modelling of densities with bounded support. This could be achieved by using the SNP densities with bounded support proposed by Kim (2007). All proposed competing risks models in this thesis focused on the CIF as the target of inference. However, competing risks models based on modelling the cause-specific hazards are also popular. One approach to apply SNP densities to cause-specific

hazards models would be to extend the SNP Cox proportional hazards model developed by Zhang & Davidian (2008) to competing risks.

6.2.4 Weighted analyses of composite endpoints

The proposed methods for the weighted comparison of composite endpoints focused on Wald-type inference and two-group comparisons. This suggests two potential areas for future research: First, it would be interesting to extend the proposed multiplicity adjustment to score- and likelihood-ratio tests which are known to have better finite-sample properties if the sample size or event rates are low. Second, extensions of the proposed test statistics to more than two groups, stratified analyses, or covariate-adjusted analyses could be pursued. Finally, the usage and relevance of weighted test statistics and associated methods for multiplicity control could be explored for more general endpoints, e.g. the joint analysis of favourable and harmful events.

Bibliography

Aalen, O. (1978), 'Nonparametric Estimation of Partial Transition Probabilities in Multiple Decrement Models', *The Annals of Statistics* **6**(3), 534–545.

Adams, R. A. & Fournier, J. J. (2003), *Sobolev spaces*, 2 edn, Academic Press.

Agresti, A. (2002), *Categorical data analysis*, 2 edn, John Wiley & Sons, Inc.

Allignol, A., Schumacher, M. & Beyersmann, J. (2011), 'Empirical Transition Matrix of Multi-State Models: The etm Package', *Journal of Statistical Software* **38**(4), 1–15.

Andersen, K., Borgan, O., Gill, R. D. & Keiding, N. (1997), *Statistical Models Based on Counting Processes*, Springer New York.

Andersen, P. K., Borgan, O., Hjort, N. L., Arjas, E., Stene, J. & Aalen, O. (1985), 'Counting Process Models for Life History Data : A Review', *Scandinavian Journal of Statistics* **12**(2), 97–158.

Andersen, P. K. & Keiding, N. (2012), 'Interpretability and importance of functionals in competing risks and multistate models.', *Statistics in medicine* **31**(11-12), 1074–88.

Arjyal, A., Basnyat, B., Koirala, S., Karkey, A., Dongol, S., Agrawaal, K. K., Shakya, N., Shrestha, K., Sharma, M., Lama, S., Shrestha, K., Khatri, N. S., Shrestha, U., Campbell, J. I., Baker, S., Farrar, J., Wolbers, M. & Dolecek, C. (2011), 'Gatifloxacin versus chloramphenicol for uncomplicated enteric fever: an open-label, randomised, controlled trial.', *The Lancet. Infectious diseases* **11**(6), 445–54.

Armstrong, P. W., Westerhout, C. M., Van de Werf, F., Califf, R. M., Welsh, R. C., Wilcox, R. G. & Bakal, J. a. (2011), 'Refining clinical trial composite outcomes: an application to the Assessment of the Safety and Efficacy of a New Thrombolytic-3 (ASSENT-3) trial.', *American heart journal* **161**(5), 848–54.

Arne Henningsen & Toomet, O. (2011), 'maxLik: A package for maximum likelihood estimation in R', *Computational Statistics* **26**(3), 443–458.

Bajorunaite, R. & Klein, J. P. (2007), 'Two-sample tests of the equality of two cumulative incidence functions', *Computational Statistics & Data Analysis* **51**(9), 4269–4281.

Bajorunaite, R. & Klein, J. P. (2008), 'Comparison of failure probabilities in the presence of competing risks', *Journal of Statistical Computation and Simulation* **78**(10), 951–966.

Bakal, J. a., Westerhout, C. M. & Armstrong, P. W. (2012), 'Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials', *Statistical Methods in Medical Research* .

Benichou, J. & Gail, M. H. (1990), 'Estimates of absolute cause-specific risk in cohort studies', *Biometrics* **46**(3), 813–826.

Beyersmann, J., Allignol, A. & Schumacher, M. (2012), *Competing Risks and Multistate Models with R*, Springer New York, New York, NY.

Beyersmann, J., Dettenkofer, M., Bertz, H. & Schumacher, M. (2007), 'A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards.', *Statistics in medicine* **26**(30), 5360–9.

Beyersmann, J., Latouche, A., Buchholz, A. & Schumacher, M. (2009), 'Simulating competing risks data in survival analysis.', *Statistics in medicine* **28**(6), 956–71.

Beyersmann, J. & Schumacher, M. (2007), 'Misspecified regression model for the subdistribution hazard of a competing risk.', *Statistics in medicine* **26**(7), 1649–51.

Beyersmann, J. & Schumacher, M. (2008), 'Time-dependent covariates in the proportional subdistribution hazards model for competing risks.', *Biostatistics (Oxford, England)* **9**(4), 765–76.

Bicanic, T., Muzoora, C., Brouwer, A. E., Meintjes, G., Longley, N., Taseera, K., Rebe, K., Loyse, A., Jarvis, J., Bekker, L.-G., Wood, R., Limmathurotsakul, D., Chierakul, W., Stepniewska, K., White, N. J., Jaffar, S. & Harrison, T. S. (2009), 'Independent association between rate of clearance of infection and clinical outcome of HIV-associated cryptococcal meningitis: analysis of a combined cohort of 262 patients.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **49**(5), 702–9.

Bierens, H. J. (2014), 'Consistency and Asymptotic Normality of Sieve ML Estimators Under Low-Level Conditions', *Econometric Theory* **30**(05), 1–56.

Birgé, L. & Massart, P. (1998), 'Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence', *Bernoulli* **4**(3), 329.

Boyd, S. & Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press, Cambridge.

Cain, K. C. & Lange, N. T. (1984), 'Approximate case influence for the proportional hazards regression model with censored data.', *Biometrics* **40**(2), 493–499.

Cannon, C. P. (1997), 'Clinical perspectives on the use of composite endpoints', *Controlled Clinical Trials* **18**(6), 517–529.

Checkley, W., Brower, R. G. & Muñoz, A. (2010), 'Inference for mutually exclusive competing events through a mixture of generalized gamma distributions.', *Epidemiology (Cambridge, Mass.)* **21**(4), 557–65.

Chen, X. (2007), 'Large sample sieve estimation of semi-nonparametric models', *Handbook of Econometrics* **6**, 5549–5632.

Chen, X. & Shen, X. (1998), 'Sieve extremum estimates for weakly dependent data', *Econometrica* **66**(2), 289–314.

Collett, D. (2003), *Modelling Survival Data in Medical Research*, 2 edn, Chapman and Hall/CRC.

Coppejans, M. & Gallant, A. (2002), 'Cross-validated SNP density estimates', *Journal of Econometrics* **110**(1), 27–65.

Cordoba, G., Schwartz, L., Woloshin, S., Bae, H. & Gø tzsche, P. C. (2010), 'Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review.', *BMJ (Clinical research ed.)* **341**, c3920.

Cox, D. (1972), 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), 187–220.

DasGupta, A., Cai, T. T. & Brown, L. D. (2001), 'Interval Estimation for a Binomial Proportion', *Statistical Science* **16**(2), 101–133.

Davison, A. C. (2008), *Statistical Models*, 1 edn, Cambridge University Press.

Day, J. N., Chau, T. T. H., Wolbers, M., Mai, P. P., Dung, N. T., Mai, N. H., Phu, N. H., Nghia, H. D., Phong, N. D., Thai, C. Q., Thai, L. H., Chuong, L. V., Sinh, D. X., Duong, V. a., Hoang, T. N., Diep, P. T., Campbell, J. I., Sieu, T. P. M., Baker, S. G., Chau, N. V. V., Hien, T. T., Lalloo, D. G. & Farrar, J. J. (2013), 'Combination antifungal therapy for cryptococcal meningitis.', *The New England journal of medicine* **368**(14), 1291–302.

Doehler, K. & Davidian, M. (2008), '"Smooth" inference for survival functions with arbitrarily censored data', *Statistics in Medicine* **27**, 5421–5439.

Eastwood, B. J. & Gallant, A. R. (1991), 'Adaptive Rules for Seminonparametric Estimators That Achieve Asymptotic Normality', *Econometric Theory* **7**(3), 307.

Escobar, L. A. & Meeker, W. Q. (1992), 'Assessing influence in regression analysis with censored data.', *Biometrics* **48**(2), 507–528.

Everitt, B. & Skrondal, A. (2007), *The Cambridge Dictionary of Statistics*, Vol. 49, 4 edn, Cambridge University Press.

Farrington, C. P. (2000), 'Residuals for proportional hazards models with interval-censored survival data.', *Biometrics* **56**(2), 473–482.

Fenton, V. M. & Gallant, A. R. (1996*a*), 'Convergence rates of SNP density estimators', *Econometrica* **64**(3), 719–727.

Fenton, V. M. & Gallant, A. R. (1996*b*), 'Qualitative and asymptotic performance of SNP density estimators', *Journal of Econometrics* **74**(1), 77–118.

Ferréira-González, I., Permanyer-Miralda, G., Busse, J. W., Bryant, D. M., Montori, V. M., Alonso-Coello, P., Walter, S. D. & Guyatt, G. H. (2007), 'Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns.', *Journal of clinical epidemiology* **60**(7), 651–7; discussion 658–62.

Fine, J. P. & Gray, R. J. (1999), 'A proportional hazards model for the subdistribution of a competing risk', *Journal of the American Statistical Association* **94**(446), 496–509.

Freemantle, N. & Calvert, M. (2007*a*), 'Composite outcomes-final comment for now...', *Journal of Clinical Epidemiology* **60**(7), 662.

Freemantle, N. & Calvert, M. (2007*b*), 'Weighing the pros and cons for composite outcomes in clinical trials', *Journal of Clinical Epidemiology* **60**(7), 658–659.

Gail, M. (1975), 'A review and critique of some models used in competing risk analysis', *Biometrics* **31**(1), 209–222.

Gail, M. H. & Pfeiffer, R. M. (2005), 'On criteria for evaluating models of absolute risk.', *Biostatistics (Oxford, England)* **6**(2), 227–39.

Gallant, A. R. & Marie Davidian (2010), 'The nonlinear mixed effects model with a smooth random effects density', *Biometrika* **80**(3), 475–488.

Gallant, A. R. & Nychka, D. W. (1987), 'Semi-Nonparametric Maximum Likelihood Estimation', *Econometrica: Journal of the Econometric Society* **55**(2), 363–390.

Gallant, A. R. & Tauchen, G. (1993), A nonparametric approach to nonlinear time series analysis: estimation and simulation, *in* 'New Directions in Time Series Analysis', Vol. 46, Springer New York, chapter 2, pp. 71–92.

Gaynor, J., Feuer, E. J. & Tan, C. C. (1993), 'On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data', *Journal of the American Statistical Association* 88(422), 400–409.

Geman, S. & Hwang, C.-R. (1982), 'Nonparametric Maximum Likelihood Estimation by the Method of Sieves', *The Annals of Statistics* **10**(2), 401–414.

Gentleman, R. & Vandal, A. C. (2002), 'Nonparametric estimation of the bivariate CDF for arbitrarily censored data', *Canadian Journal of Statistics* **30**(4), 557–571.

Geskus, R. B. (2011), 'Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring.', *Biometrics* **67**(1), 39–49.

Glasziou, P. P., Simes, R. J. & Gelber, R. D. (1990), 'Quality adjusted survival analysis.', *Statistics in medicine* 9(11), 1259–1276.

Gooley, T. A., Leisenring, W., Crowley, J. & Storer, B. E. (1999), 'Estimation of failure probabilities in the presence of competing risks: new representations of old estimators', *Statistics in medicine* **18**(16), 695–706.

Grambauer, N., Schumacher, M. & Beyersmann, J. (2010), 'Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified.', *Statistics in medicine* **29**(7-8), 875–84.

Gray, B. (2013), 'cmprsk: Subdistribution Analysis of Competing Risks'.
URL: *http://cran.r-project.org/package=cmprsk*

Gray, R. (1988), 'A class of K-sample tests for comparing the cumulative incidence of a competing risk', *The Annals of statistics* **16**(3), 1141–1154.

Grenander, U. (1981), *Abstract Inference*, 1 edn, John Wiley & Sons, Inc.

Groeneboom, P., Maathuis, M. H. & Wellner, J. a. (2008*a*), 'Current status data with competing risks: Consistency and rates of convergence of the MLE', *The Annals of Statistics* **36**(3), 1031–1063.

Groeneboom, P., Maathuis, M. H. & Wellner, J. a. (2008*b*), 'Current status data with competing risks: Limiting distribution of the MLE', *Annals of Statistics* **36**(3), 1064–1089.

Heckman, J. J. & Honoré, B. E. (1989), 'The identifiability of the competing risks model', *Biometrika* **76**(2), 325–330.

Höfler, M. (2005), 'Causal inference based on counterfactuals.', *BMC medical research methodology* **5**, 28.

Hollander, M. & Proschan, F. (1979), 'Testing to determine the underlying distribution using randomly censored data', *Biometrics* **35**(2), 393–401.

Holt, J. D. (1978), 'Competing Risk Analyses with Special Reference to Matched Pair Experiments', *Biometrika* **65**(1), 159–165.

Hong, K.-S., Ali, L. K., Selco, S. L., Fonarow, G. C. & Saver, J. L. (2011), 'Weighting components of composite end points in clinical trials: an approach using disability-adjusted life-years.', *Stroke; a journal of cerebral circulation* **42**(6), 1722–9.

Honoré, B. E. & Lleras-Muney, A. (2006), 'Bounds in Competing Risks Models and the War on Cancer', *Econometrica* **74**(6), 1675–1698.

Hudgens, M. G., Li, C. & Fine, J. P. (2014), 'Parametric likelihood inference for interval censored competing risks data.', *Biometrics* **70**(1), 1–9.

Hudgens, M. G., Satten, G. A. & Longini, I. M. (2001), 'Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation.', *Biometrics* **57**(1), 74–80.

Jeong, J.-H. & Fine, J. P. (2007), 'Parametric regression on cumulative incidence function', *Biostatistics* **8**(2), 184–196.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, 2 edn, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Kauermann, G. & Carroll, R. J. (2001), 'A Note on the Efficiency of Sandwich Covariance Matrix Estimation', *Journal of the American Statistical Association* **96**(456), 1387–1396.

Kim, K. I. (2007), 'Uniform convergence rate of the seminonparametric density estimator and testing for similarity of two unknown densities', *The Econometrics Journal* **10**(1), 1–34.

Koller, M. T., Raatz, H., Steyerberg, E. W. & Wolbers, M. (2012), 'Competing risks and the clinical community: irrelevance or ignorance?', *Statistics in medicine* **31**(11-12), 1089–97.

Kooperberg, C. & Stone, C. J. (1992), Logspline Density Estimation for Censored Data, Technical Report 226.

Krailo, M. D. & Pike, M. C. (1983), 'Estimation of the distribution of age at natural menopause from prevalence data.', *American journal of epidemiology* **117**(3), 356–361.

Kudo, A. (1963), 'A Multivariate Analogue of the One-Sided Test', *Biometrika* **50**(3/4), 403–418.

Kuk, A. & Chen, C.-h. (1992), 'A mixture model combining logistic regression with proportional hazards regression', *Biometrika* **3**.

Larson, M. & Dinse, G. (1985), 'A mixture model for the regression analysis of competing risks data', *Applied Statistics* **34**(3), 201–211.

Latouche, A., Boisson, V., Chevret, S. & Porcher, R. (2007), 'Misspecified regression model for the subdistribution hazard of a competing risk.', *Statistics in medicine* **26**(5), 965–74.

Lau, B., Cole, S. R. & Gange, S. J. (2011), 'Parametric mixture models to evaluate and summarize hazard ratios in the presence of competing risks with time-dependent hazards and delayed entry.', *Statistics in medicine* **30**(6), 654–65.

Lau, B., Cole, S. R., Moore, R. D. & Gange, S. J. (2008), 'Evaluating competing adverse and beneficial outcomes using a mixture model.', *Statistics in medicine* **27**(21), 4313–27.

Lefkopoulou, M. & Ryan, L. (1993), 'Global tests for multiple binary outcomes', *Biometrics* **49**(4), 975–988.

Li, C. & Fine, J. P. (2013), 'Smoothed nonparametric estimation for current status competing risks data', *Biometrika* **100**(1), 173–187.

Lim, E., Brown, A., Helmy, A., Mussa, S. & Altman, D. G. (2008), 'Composite outcomes in cardiovascular research: A survey of randomized trials', *Annals of Internal Medicine* **149**(9), 612–617.

Lin, D. Y. (1997), 'Non-parametric inference for cumulative incidence functions in competing risks studies', *Statistics in Medicine* **16**(8), 901–910.

Maathuis, M. (2013), 'MLEcens: Computation of the MLE for bivariate (interval) censored data'. **URL:** *http://cran.r-project.org/package=MLEcens*

Maathuis, M. H. (2003), Nonparametric maximum likelihood estimation for bivariate censored data, Master's thesis, Delft University of Technology.

Maathuis, M. H. (2006), Nonparametric estimation for current status data with competing risks, PhD thesis, University of Washington.

Maller, R. A. & Zhou, X. (2002), 'Analysis of parametric models for competing risks', *Statistica Sinica* **12**, 725–750.

Mascha, E. J. & Imrey, P. B. (2010), 'Factors affecting power of tests for multiple binary outcomes.', *Statistics in medicine* **29**(28), 2890–904.

Mascha, E. J. & Sessler, D. I. (2011), 'Statistical grand rounds: design and analysis of studies with binary- event composite endpoints: guidelines for anesthesia research.', *Anesthesia and analgesia* **112**(6), 1461–71.

Mascha, E. J. & Turan, A. (2012), 'Joint hypothesis testing and gatekeeping procedures for studies with multiple endpoints.', *Anesthesia and analgesia* **114**(6), 1304–17.

McCall, B. P. (1996), 'The Identifiability of the Mixed Proportional Hazards Model with Time-Varying Coefficients', *Econometric Theory* **12**(3), 733.

Mell, L. & Jeong, J. (2010), 'Pitfalls of using composite primary end points in the presence of competing risks', *Journal of Clinical Oncology* **28**(28), 4297–4299.

Ng, A. S. K. & McLachlan, G. J. (2003), 'An EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data', *Statistics in Medicine* **22**(7), 1097–1111.

Nicolaie, M. A., van Houwelingen, H. C. & Putter, H. (2010), 'Vertical modeling: A pattern mixture approach for competing risks modeling', *Statistics in medicine* **29**(11), 1190–1205.

Pandit, A., Arjyal, A., Day, J. N., Paudyal, B., Dangol, S., Zimmerman, M. D., Yadav, B., Stepniewska, K., Campbell, J. I., Dolecek, C., Farrar, J. J. & Basnyat, B. (2007), 'An open randomized comparison of gatifloxacin versus cefixime for the treatment of uncomplicated enteric fever', *PLoS ONE* **2**(6), e542.

Pepe, M. S. (1991), 'Inference for events with dependent risks in multiple endpoint studies', *Journal of the American Statistical Association* **86**(415), 770–778.

Pepe, M. S. & Fleming, T. (1989), 'Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data.', *Biometrics* **45**(2), 497–507.

Pepe, M. S. & Fleming, T. R. (1991), 'Weighted Kaplan-Meier statistics: Large sample and optimality considerations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **53**(2), 341–352.

Pepe, M. S. & Mori, M. (1993), 'Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data?', *Statistics in medicine* **12**(8), 737–751.

Pettitt, A. & Stephens, M. (1976), 'Modified Cramer-von Mises Statistics for Censored Data', *Biometrika* **63**(2), 291–298.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Jr., Flournoy, N., Farewell, V. T. & Breslow, N. E. (1978), 'The Analysis of Failure Times in the Presence of Competing Risks.', *Biometrics* **34**(4), 541–544.

Putter, H., Fiocco, M. & Gekus, R. B. (2007), 'Tutorial in biostatistics: Competing risk and multi-state models', *Statistics in Medicine* **26**(11), 2389–2430.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.r-project.org/*

Rauch, G. & Beyersmann, J. (2013), 'Planning and evaluating clinical trials with composite time-to-first-event endpoints in a competing risk framework.', *Statistics in medicine* **32**(21), 3595–608.

Sampson, U. K. a., Metcalfe, C., Pfeffer, M. a., Solomon, S. D. & Zou, K. H. (2010), 'Composite outcomes: weighting component events according to severity assisted interpretation but reduced statistical power.', *Journal of clinical epidemiology* **63**(10), 1156–8.

Scheffe, H. (1953), 'A Method for Judging all Contrasts in the Analysis of Variance', *Biometrika* **40**(1), 87–104.

Schumacher, M. (1984), 'Two-Sample Tests of Cramér–von Mises-and Kolmogorov–Smirnov-Type for Randomly Censored Data', *International Statistical Review / Revue Internationale de Statistique* **52**(3), 263–281.

Shapiro, A. (1988), 'Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis', *International Statistical Review / Revue Internationale de Statistique* **56**(1), pp. 49–62.

Shapiro, A. (2003), 'Scheffe's method for constructing simultaneous confidence intervals subject to cone constraints', *Statistics & Probability Letters* **64**(4), 403–406.

Shen, X. (1997), 'On methods of sieves and penalization', *The Annals of Statistics* **25**(6), 2555–2591.

Therneau, T. M. & Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Thwaites, G. E., Thwaites, G. E., Bang, N. D., Bang, N. D., Dung, N. H., Dung, N. H., Quy, H. T., Quy, H. T., Oanh, D. T. T., Oanh, D. T. T., Thoa, N. T. C., Thoa, N. T. C., Hien, N. Q., Hien, N. Q., Thuc, N. T., Thuc, N. T., Hai, N. N., Hai, N. N., Lan, N. T. N., Lan, N. T. N., Lan, N. N., Lan, N. N., Duc, N. H., Duc, N. H., Tuan, V. N., Tuan, V. N., Hiep, C. H., Hiep, C. H., Chau, T. T. H., Chau, T. T. H., Mai, P. P., Mai, P. P., Dung, N. T., Dung, N. T., Stepniewska, K., Stepniewska, K., White, N. J., White, N. J., Hien, T. T., Hien, T. T. & Farrar, J. J. (2004), 'Dexamethasone for the Treatment of Tuberculous Meningitis in Adolescents and Adults', *The New England journal of medicine* **351**, 1741–1751.

Titman, A. C. & Sharples, L. D. (2010), 'Model diagnostics for multi-state models.', *Statistical methods in medical research* **19**, 621–651.

Tong, B. C., Huber, J. C., Ascheim, D. D., Puskas, J. D., Jr, B. F., Blackstone, E. H. & Smith, P. K. (2012), 'Weighting Composite Endpoints in Clinical Trials: Essential Evidence for the Heart Team', **94**(6), 1908–1913.

Török, M. E., Yen, N. T. B., Chau, T. T. H., Mai, N. T. H., Phu, N. H., Mai, P. P., Dung, N. T., Chau, N. V. V., Bang, N. D., Tien, N. A., Minh, N. H., Hien, N. Q., Thai, P. V. K., Dong, D. T., Anh, D. T. T., Thoa, N. T. C., Hai, N. N., Lan, N. N., Lan, N. T. N., Quy, H. T., Dung, N. H., Hien, T. T., Chinh, N. T., Simmons, C. P., de Jong, M., Wolbers, M. & Farrar, J. J. (2011), 'Timing of initiation of antiretroviral therapy in human immunodeficiency virus (HIV)–associated tuberculous meningitis.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **52**(11), 1374–83.

Tsiatis, A. (1975), 'A nonidentifiability aspect of the problem of competing risks.', *Proceedings of the National Academy of Sciences* **72**(1), 20–22.

Van de Geer, S. (1993), 'Hellinger-consistency of certain nonparametric maximum likelihood estimators', *The Annals of Statistics* **21**(1), 14–44.

Van der Vaart, A. (2000), *Asymptotic statistics*, 1 edn, Cambridge University Press.

Volinsky, C. T. & Raftery, A. E. (2000), 'Bayesian information criterion for censored survival models.', *Biometrics* **56**, 256–262.

Wolbers, M., Koller, M. T., Stel, V. S., Schaer, B., Jager, K. J., Leffondre, K. & Heinze, G. (2014), 'Competing risks analyses: objectives and approaches.', *European heart journal* **35**(42), 2936–41.

Wolbers, M., Koller, M. T., Witteman, J. C. M. & Steyerberg, E. W. (2009), 'Prognostic models with competing risks methods and application to coronary risk prediction', *Epidemiology* **20**(4), 555–561.

Wong, W. & Shen, X. (1995), 'Probability inequalities for likelihood ratios and convergence rates of sieve MLEs', *The Annals of Statistics* **23**(2), 339–362.

Zhang, D. & Davidian, M. (2010), 'Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data Linear Mixed Models with Flexible Distributions Random Effects for Longitudinal Data', *Society* **57**(3), 795–802.

Zhang, M. & Davidian, M. (2008), '"Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data.', *Biometrics* **64**(2), 567–76.

Appendix A

Appendix

Throughout this appendix, the following notation is used (in addition to the notation introduced in Chapters 2 and 3):

- $P_K(z)$ denotes a polynomial on $\mathbb{R}$ of degree K, $P_K(z, \mathbf{a})$ also refers to a polynomial of degree K but with specific coefficients $\mathbf{a} = (a_0, a_1, \ldots, a_K)^t$ i.e. $P_K(z, \mathbf{a}) = \sum_{i=0}^{K} a_i z^i$
- $\varphi(z)$ and Φ denote the density and the cumulative distribution function of the standard normal distribution respectively
- $\varepsilon(z)$ denotes the density of the exponential distribution with rate equal to 1
- $\psi(z)$ denotes a (base) density e.g. $\varphi(z)$ or $\varepsilon(z)$
- $h_K(z; \mathbf{a})$ denotes a SNP density with polynomial $P_K(z, \mathbf{a})$ i.e. $h_k(z; \mathbf{a}) = P_K^2(z, \mathbf{a})\psi(z)$

A.1 Application of Lemma A.5 of Gallant and Nychka

Lemma: *Let* $\mu(z) = (1 + z^2)^{\delta}$ *for some* $\delta > 0$*; and for some* $\mathcal{B} > 0$ *let*

$$\mathcal{V} = \left\{ f : \; f(z) = P_K(z)\sqrt{\varphi(z)}; \; \left\| P_K(z)\sqrt{\varphi(z)} \right\|_{m,2,\mu} < \mathcal{B}, \; K = 0, 1, 2, \ldots \right\}$$

where $\varphi(z)$ *is a density function such that* $\left(\int \sqrt{\varphi(u)} du \right)^{-1} \sqrt{\varphi(z)}$ *is a strictly positive density with a moment generating function. Then* $\mathcal{V}$ *is dense in* $W_{m,2,\mu}$.

Proof: First let $D = \int \sqrt{\varphi(z)} dz$. Because $D^{-1}\sqrt{\varphi(z)}$ is a strictly positive density function with a moment generating function,

$$\mathcal{V}' = \left\{ f : \; f(z) = P_K(z) D^{-1} \sqrt{\varphi(z)}; \; K = 0, 1, 2, \ldots \right\} = D^{-1} \times \mathcal{V}$$

is dense in $W_{m,2,\mu}$ according to Lemma A.5 in Gallant & Nychka (1987), and so is $\mathcal{V}$ *Q.E.D.*

A.2 Likelihood calculations for fixed SNP polynomial degrees

This section reviews material presented in the Web appendix of Zhang & Davidian (2008) and details how to carry out the log-likelihood calculations in Sections 3.2 and 4.2 were carried out.

A.2.1 Spherical coordinates

Let $\mathbf{W}=(1, Z, Z^2, \ldots, Z^K)^t$ be the random vector whose Z has density $\psi(z)$ and $\mathbf{w}=(1, z, z^2, \ldots, z^K)^t$. Then $h_k(z;\mathbf{a}) = \left(\sum_{i=0}^{K} a_i z^i\right)^2 \psi(z)$ can be rewritten as $\mathbf{a}^t\mathbf{w}\mathbf{w}^t\mathbf{a}\psi(z)$. Thus, $\int h_k(z;\mathbf{a})dz = 1$ means $\int \mathbf{a}^t\mathbf{w}\mathbf{w}^t\mathbf{a}\psi(z)dz = 1$ or $\mathbf{a}^tA\mathbf{a} = 1$, where $A = \int \mathbf{w}\mathbf{w}^t\psi(z)dz$. For example when $K = 2$

$$\mathbf{w}\mathbf{w}^t = \begin{pmatrix} 1 & z & z^2 \\ z & z^2 & z^3 \\ z^2 & z^3 & z^4 \end{pmatrix}$$

and

$$A = \begin{pmatrix} 1 & \int z\psi(z)dz & \int z^2\psi(z)dz \\ \int z\psi(z)dz & \int z^2\psi(z)dz & \int z^3\psi(z)dz \\ \int z^2\psi(z)dz & \int z^3\psi(z)dz & \int z^4\psi(z)dz \end{pmatrix} = \begin{pmatrix} 1 & E_\psi(Z) & E_\psi(Z^2) \\ E_\psi(Z) & E_\psi(Z^2) & E_\psi(Z^3) \\ E_\psi(Z^2) & E_\psi(Z^3) & E_\psi(Z^4) \end{pmatrix}$$

For $\psi(z)$ being the standard normal density function $\varphi(z)$ or the exponential distribution with rate 1 $\varepsilon(z)$, A is known and positive definite. Specifically, for $\psi(z) = \varphi(z)$, $E_\psi(Z^i) = 0$ for odd i, and $E_\psi(Z^i) = \frac{i!}{2^{i/2}(i/2)!}$ for even i; and for $\psi(z) = \varepsilon(z)$ $E_\psi(Z^i) = i!$. As A is positive definite, it can be decomposed as $A = B^tB$ for some positive definite matrix B e.g. by using Cholesky's decomposition. Thus, $\mathbf{a}^tA\mathbf{a} = \mathbf{a}^tB^tB\mathbf{a}$, and if $C(K) = B\mathbf{a} = (c_1, \ldots, c_{K+1})^t$ then $C(K)^tC(K) = 1$. Hence, $C(K)$ lies on the unit sphere suggesting the spherical parametrization

$$\begin{aligned} c_1 &= \sin(\phi_1) \\ c_2 &= \cos(\phi_1)\sin(\phi_2) \\ \cdots &\quad \cdots \\ c_K &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\sin(\phi_K) \\ c_{K+1} &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\cos(\phi_K) \end{aligned}$$

where $\phi_K = (\phi_1, \phi_2, \ldots, \phi_K)^t$ are the spherical coordinates with $-\pi/2 < \phi_j \leq \pi/2, j = 1, \ldots, K$. Of note it is only required that the space of the ϕ spans one half of the unit sphere as $h_k(z; \mathbf{a}) = \mathbf{a}^t \mathbf{w}\mathbf{w}^t \mathbf{a}\psi(z)$ are the same as $h_k(z; -\mathbf{a})$; and hence, the corresponding $C(K)$ and $-C(K)$ also specify the same $h_k(z; \mathbf{a})$. Of note, the spherical parametrization in the Web appendix of Zhang & Davidian (2008) contains some typos (the specification of c_K and the given range for ϕ_K are are wrong) but they are correctly reported as above in the main text of that paper.

From now on, I use $h_k(z; \phi_K)$ and $P_K(z; \phi_K)$ to refer to the SNP density and its polynomial whose coefficients are specified by the spherical coordinates ϕ_K.

A.2.2 SNP densities

In this section, I lay down the specific formulas used in the calculation of the density of a random survival T_0 specified by $\log T_0 = \mu + \sigma Z$, $\mu \in \mathbb{R}$, $\sigma > 0$, with Z having SNP distribution of degree K with a standard normal base density or $\exp(Z)$ following a SNP distribution of degree K with an exponential base density. For convenience, I refer to the former as the normal case and the latter as the exponential case. In both cases, let the SNP density be specified by the spherical coordinates ϕ_K. Thus, the parameter vector of the whole model for T_0 is $\theta = (\mu, \sigma, \phi_K)$. According to Zhang & Davidian (2008), for the normal case, the density of T_0 is:

$$f_{0,K}(t; \theta) = (t\sigma)^{-1} P_K^2 \left(\frac{\log t - \mu}{\sigma}; \phi_K \right) \varphi \left(\frac{\log t - \mu}{\sigma} \right)$$

while for the exponential case it is

$$f_{0,K}(t; \theta) = \left(\sigma e^{\mu/\sigma} \right)^{-1} t^{(1/\sigma - 1)} P_K^2 \left\{ (t/e^{\mu})^{1/\sigma}; \phi \right\} \varepsilon \left\{ (t/e^{\mu}) \right\}$$

Under the AFT model, where $\log T = \mathbf{X}^t \beta + \log T_0 = \mathbf{X}^t \beta + \mu + \sigma Z$ with β and $\mathbf{X}$ being the regression coefficient and covariate vector, respectively, the density of T is given by

$$f_K(t; \theta, \beta, \mathbf{X}) = \exp(-\mathbf{X}^t \beta) \left(e^{-\mathbf{X}^t \beta} t\sigma \right)^{-1} P_K^2 \left(\frac{\log t - \mathbf{X}^t \beta - \mu}{\sigma}; \phi_K \right) \varphi \left(\frac{\log t - \mathbf{X}^t \beta - \mu}{\sigma} \right)$$

and

$$f_K(t; \theta, \beta, \mathbf{X}) = \exp(-\mathbf{X}^t \beta) \left(\sigma e^{\mu/\sigma} \right)^{-1} \left(e^{-\mathbf{X}^t \beta} t \right)^{(1/\sigma - 1)} P_K^2 \left\{ \left(e^{-\mathbf{X}^t \beta} t / e^{\mu} \right)^{1/\sigma}; \phi \right\} \varepsilon \left\{ \left(e^{-\mathbf{X}^t \beta} t / e^{\mu} \right) \right\}$$

respectively.

A.2.3 SNP survival function

Under the same set-up of the previous subsection, in general the survival function of T_0 is

$$S_{0,K}(t;\theta) = \int_c^\infty P_K^2(z;\phi_K)\,\psi(z)dz$$

the lower bound c is $\frac{\log t-\mu}{\sigma}$ in the normal case and $(t/e^\mu)^{1/\sigma}$ in the exponential case. In both cases, $P_K^2(z;\phi_K)$ can be expanded to a polynomial of degree $2K$: $\sum_{k=0}^{2K} d_k z^k$; hence $S_{0,K}(t;\theta) = \sum_{k=0}^{2K} d_k \left(\int_c^\infty z^k\psi(z)dz\right)$. Let $I(k,c) = \int_c^\infty z^k\psi(z)dz$, $k = 0,1,\ldots,2K$ and a recursion formula can be derived to compute $I(k,c)$ using integration by parts.

Specifically in the normal case, $I(k,c) = c^{k-1}\varphi(c) + (k-1)I(k-2,c)$ for $k \geq 2$, and $I(0,c) = 1-\Phi(c)$, $I(1,c) = \varphi(c)$. For the exponential case, $I(k,c) = c^k\varepsilon(c) + kI(k-1,c)$ for $k > 0$, and $I(0,c) = e^{-c}$.

Under the AFT model, where $\log T = \mathbf{X}^t\beta + \mu + \sigma Z$, the survival function becomes

$$S_K(t;\theta,\beta,\mathbf{X}) = S_{0,K}\left(e^{-\mathbf{X}^t\beta}t;\theta\right)$$

A.3 Moment calculations

Equation (3.3.2) in Subsection 3.3.2 requires the moment of Z in the normal and exponential cases. For the normal case, after expanding the SNP polynomial as in Subsection A.2.3, it is easy to see that

$$\begin{aligned} E(Z) &= \textstyle\sum_{i=0}^{2K} d_i E_\varphi\left(Z^{i+1}\right) \\ E\left(Z^2\right) &= \textstyle\sum_{i=0}^{2K} d_i E_\varphi\left(Z^{i+2}\right) \end{aligned}$$

where $E_\varphi\left(Z^i\right)$ are the i^{th} moment of the standard normal distribution which are well known and were explicitly given in Appendix A.2.1.

In the exponential case, we need to calculate $E(Z)$ and $E\left(Z^2\right)$ if $Z^* = \exp(Z)$ has a SNP density of degree K given by $f_{Z^*}(z) = h_K(z) = \left(\sum_{i=0}^K a_i z^i\right)^2 e^{-z} = \sum_{i=0}^{2K} d_i z^i e^{-z}$. For convenience, I calculate the expectation of $Y = -\log Z^* = -Z$ first. Using a change of variable, we have

$$f_Y(y) = \left(\sum_{i=0}^{2K} d_i e^{-yi}\right)\exp\left(-y-e^{-y}\right)$$

Therefore,

$$\begin{aligned} E(Y) &= \textstyle\int_{-\infty}^\infty y\left(\sum_{i=0}^{2K} d_i e^{-yi}\right)\exp\left(-y-e^{-y}\right)dy \\ &= \textstyle\sum_{i=0}^{2K} d_i \int_{-\infty}^\infty y e^{-yi}\exp\left(-y-e^{-y}\right)dy \end{aligned}$$

Let $I(n) = \int_{-\infty}^{\infty} ye^{-yn} \exp\left(-y - e^{-y}\right) dy,\ n = 0, 1, 2, \ldots, 2K$. We have

$$I(0) = \int_{-\infty}^{\infty} y \exp\left(-y - e^{-y}\right) dy = \gamma$$

Here γ is the Euler constant. For $n \geq 1$, using integration by parts, we have

$$I(n) = nI(n-1) - \int_{-\infty}^{\infty} \exp\left(-yn - e^{-y}\right) dy$$

where $\int_{-\infty}^{\infty} \exp\left(-yn - e^{-y}\right) dy = 1$ for $n = 0$ as $\exp\left(-e^{-y}\right)$ is the density of the standard Gumbel distribution, and again, using integration by parts yields

$$\int_{-\infty}^{\infty} \exp\left(-yn - e^{-y}\right) dy = (n-1) \int_{-\infty}^{\infty} \exp\left(-y(n-1) - e^{-y}\right) dy$$

for $n \geq 1$ and thus $\int_{-\infty}^{\infty} \exp\left(-yn - e^{-y}\right) dy = (n-1)!$. Hence,

$$I(n) = nI(n-1) - (n-1)!\,.$$

Thus, $E(Y) = \sum_{i=0}^{2K} d_i I(n)$ with $I(0) = \gamma$ and $I(n) = nI(n-1) - (n-1)!,\ n = 1, 2, .., 2K$. Finally, $E(Z) = -E(Y)$.

Now, we calculate $E\left(Z^2\right) = E\left(Y^2\right)$ which is

$$E\left(Y^2\right) = \sum_{i=0}^{2K} d_i \int_{-\infty}^{\infty} y^2 e^{-yi} \exp\left(-y - e^{-y}\right) dy$$

Let $H(n) = \int_{-\infty}^{\infty} y^2 e^{-yn} \exp\left(-y - e^{-y}\right) dy,\ n = 0, 1, 2, \ldots, 2K$. We have

$$H(0) = \int_{-\infty}^{\infty} y^2 \exp\left(-y - e^{-y}\right) dy = \beta$$

where β can be calculated using the moment generating function of the standard Gumbel distribution. For $n \geq 1$, using integration by parts, we have $H(n) = -2I(n-1) + nH(n-1)$.

To sum up, we have $E(\log Z) = -\sum_{i=0}^{2K} d_i I(n)$ and $E\left\{(\log Z)^2\right\} = \sum_{i=0}^{2K} d_i H(n)$ with

$$\begin{aligned}
I(0) &= \gamma \\
H(0) &= \beta \\
I(n) &= nI(n-1) - (n-1)!,\ n \geq 1 \\
H(n) &= -2I(n-1) + nH(n-1),\ n \geq 1
\end{aligned}$$

A.4 Simulation of univariate SNP random variables

This section gives a univariate interpretation of Section 3 in Gallant & Tauchen (1993) which shows how to simulate a random variable having density $h(z) = \left(\sum_{i=0}^{K} a_i z^i\right)^2 \psi(z)$. The rejection method requires an upper envelope $b(z)$ with $h(z) \leq b(z)$. To define this envelope, first expand the polynomial $\left(\sum_{i=0}^{K} a_i z^i\right)^2$ to an explicit polynomial of degree $2K$ of the form $\sum_{i=0}^{2K} d_i z^i$. The envelope is then defined as $b(z) = \sum_{i=0}^{2K} |d_i|\,|z|^i\,\psi(z)$.

A.4.1 Review of the rejection method

The rejection method for sampling from a (multivariate) density $h(z)$ depends on a non-negative, integrable function $b(z)$ that dominates $h(z)$ viz

$$0 \leq h(z) \leq b(z)$$

The domination function $b(z)$ is called an upper envelope for $h(z)$ or majorizing function. Derive a density $g(v)$ from $b(v)$ by putting

$$g(v) = b(v)\left(\int b(s)ds\right)^{-1}$$

Using $b(z)$ and $g(z)$, a sample from $h(z)$ is generated as follows. Generate the pair (u, v) by generating v from $g(v)$ and u from the uniform distribution on $[0, 1]$. If

$$u > h(v)/b(v)$$

reject the pair (u, v) and try again. Otherwise, accept $z = v$ as a sample from $h(z)$.

A.4.2 Simulation for standard normal base density

Define the envelope function as $b(z) = \left(\sum_{i=0}^{2K} |d_i|\,|z|^i\right)\varphi(z) = \frac{1}{\sqrt{2\pi}}\sum_{i=0}^{2K} |d_i|\,\frac{\Gamma\left(\frac{i+1}{2}\right)}{2^{\frac{1-i}{2}}}\chi(i+1, |z|)$. This is a weighted sum of chi-densities $\chi(i+1, |z|) = \frac{2^{\frac{1-i}{2}}}{\Gamma\left(\frac{i+1}{2}\right)}|z|^i e^{-z^2/2}$ with $i+1$ degrees of freedom. The chi-density describes the distribution of the square root of a random variable with a chi-squared distribution. The weights are then $u_i = \frac{1}{\sqrt{2\pi}}|d_i|\,\frac{\Gamma\left(\frac{i+1}{2}\right)}{2^{\frac{1-i}{2}}}$, $i = \overline{0, 2K}$. Next we get normalized weights as $w_i = \frac{u_i}{\sum_{j=0}^{2K} u_j}$. Then the simulation proceeds as follows:

Step 1: Use the multinomial distribution with cell probabilities w_i's to randomly select a number in $\{1, ..., 2K+1\}$ as the degree of freedom.

Step 2: Get a chi-distributed number by taking the square-root of the corresponding chi-square distributed number having the above degree of freedom. Call this number z.

Step 3: Randomly set the sign of z with probability of being negative equal to 0.5. Denote the new number by v.

Step 4: Generate u from the uniform distribution on $[0, 1]$. If $u > h(v)/b(v)$ return to step 1. Otherwise, accept v.

A.4.3 Simulation for standard exponential base density

Define the envelope function as $b(z) = \left(\sum_{i=0}^{2K} |d_i| z^i\right) \exp(-x) = \sum_{i=0}^{2K} |d_i| \Gamma(i+1) G(1, i+1, z)$. This is a weighted sum of gamma densities G whose shape and scale parameters are, respectively, 1 and $i + 1$. The weights are then $u_i = |d_i| \Gamma(i + 1),\ i = \overline{0, 2K}$. Next we get the normalized weights as $w_i = \frac{u_i}{\sum_{j=0}^{2K} u_j}$. Then the simulation proceeds as for the normal base density (Section A.4.2), except that in step 2 we generate a Gamma-distributed random variable.

Appendix B

Appendix

B.1 Proof for asymptotic normality of $C_{t_{max}}$

To fix ideas, assume that we are interested in the first competing risk and without loss of generality, for simplicity I combine all competing events other than the first event type to form a "second" competing event. Let $\widehat{CIF}_1(\cdot)$ and $\widehat{CIF}_2(\cdot)$ be the nonparametric estimator of, respectively $CIF_1(\cdot)$ and $CIF_2(\cdot)$; and $CIF_{1\theta}(\cdot)$ be the proposed parametric model for the true $CIF_1(\cdot)$. In addition, for a sample of n observations, the observed right-censored data are i.i.d. pairs $\left\{\tilde{T}_i = \min(T_i, C_i), \delta_i\right\}$, $i = 1, ..., n$, where T_i and C_i are the total survival time and censoring time for subject i, respectively, and δ_i is the corresponding event-status indicator. Define $Y(t) = \sum_{i=1}^{n} Y_i(t)$, $N_j(t) = \sum_{i=1}^{n} N_{ji}(t)$ and $M_j(t) = \sum_{i=1}^{n} M_{ji}(t)$, where $Y_i(t) = \mathbf{I}\left(\tilde{T}_i \geq t\right)$, $N_{ji}(t) = \mathbf{I}\left(\tilde{T}_i \leq t, \delta_j = j\right)$ and $M_{ji}(t) = N_{ji}(t) - \int_0^t Y_i(u)\lambda_j(u)du$ with $\lambda_j(\cdot)$ being the cause-specific hazard of event type j. Let $\Lambda_j(t) = \int_0^t \lambda_j(u)du$, the cumulative cause-specific hazard for event type j whose corresponding nonparametric Aalen-Nelson estimator is $A_j(t) = \int_0^t Y^{-1}(u)dN_j(u)$. The nonparametric estimates of $M_j(t)$ are $\hat{M}_j(t) = \sum_{i=1}^{n} \hat{M}_{ji}(t)$, with $\hat{M}_{ji}(t) = N_{ji}(t) - A_j(t)$. From Section 6 of Gray (1988), $M_{ji}(t)$, $j = 1, 2$; $i = 1, ..., n$ are orthogonal martingales. Then consider

$$C_{t_{max}} = CIF_{1\theta}^{-2}(t_{max}) \int_0^{t_{max}} \widehat{CIF}_1(u) dCIF_{1\theta}(u)$$

Next let

$$\begin{aligned} I_{t_{max}} &= \sqrt{n}\left\{C_{t_{max}} - CIF_{1\theta}^{-2}(t_{max}) \int_0^{t_{max}} CIF_1(u) dCIF_{1\theta}(u)\right\} \\ &= CIF_{1\theta}^{-2}(t_{max}) \int_0^{t_{max}} \sqrt{n}\left\{\widehat{CIF}_1(u) - CIF_1(u)\right\} dCIF_{1\theta}(u) \end{aligned}$$

From Lin (1997), we have

$$\sqrt{n}\left\{\widehat{CIF}_1(t) - CIF_1(t)\right\} = \sqrt{n}\left\{\int_0^t \frac{1-CIF_2(u)}{Y(u)} dM_1(u) + \int_0^t \frac{CIF_1(u)}{Y(u)} dM_2(u) - CIF_1(t)\int_0^t \frac{dM_1(u)+dM_2(u)}{Y(u)}\right\} + o_P(1)$$

where all integrals can be considered as Lebesgue-Stieltjes integrals which was informally discussed in Appendix B.2 of Kalbfleisch & Prentice (2002), and the integrands having $Y(u)$ are taken to be 0 if $Y(u) = 0$. Accordingly,

$$I_{t_{\max}} \approx \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \times \left\{\int_0^{t_{\max}} \left[\int_0^u \frac{1-CIF_2(s)}{Y(s)} dM_1(s)\right] dCIF_{1\theta}(u) + \int_0^{t_{\max}} \left[\int_0^u \frac{CIF_1(s)}{Y(s)} dM_2(s)\right] dCIF_{1\theta}(u) - \int_0^{t_{\max}} CIF_1(u)\left[\int_0^u \frac{dM_1(s)+dM_2(s)}{Y(s)}\right] dCIF_{1\theta}(u)\right\}$$

Following Bajorunaite & Klein (2007), after changing the order of integration (or using integration by parts), we have

$$\begin{aligned} I_{t_{\max}} \approx & \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \\ & \times \left\{\int_0^{t_{\max}} \left[\int_s^{t_{\max}} \frac{1-CIF_2(s)}{Y(s)} dCIF_{1\theta}(u)\right] dM_1(s) + \int_0^{t_{\max}} \left[\int_s^{t_{\max}} \frac{CIF_1(s)}{Y(s)} dCIF_{1\theta}(u)\right] dM_2(s) \right. \\ & \left. - \int_0^{t_{\max}} \left[\int_s^{t_{\max}} \frac{CIF_1(u)}{Y(s)} dCIF_{1\theta}(u)\right] [dM_1(s) + dM_2(s)]\right\} \\ = & \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \\ & \times \left\{\int_0^{t_{\max}} \left[\frac{1-CIF_2(s)}{Y(s)} \int_s^{t_{\max}} dCIF_{1\theta}(u) - \frac{1}{Y(s)} \int_s^{t_{\max}} CIF_1(u) dCIF_{1\theta}(u)\right] dM_1(s) \right. \\ & \left. + \int_0^{t_{\max}} \left[\frac{CIF_1(s)}{Y(s)} \int_s^{t_{\max}} dCIF_{1\theta}(u) - \frac{1}{Y(s)} \int_s^{t_{\max}} CIF_1(u) dCIF_{1\theta}(u)\right] dM_2(s)\right\} \end{aligned}$$

As seen $I_{t_{\max}}$ can be approximated by a sum of stochastic integrals involving predictable processes and orthogonal martingales. Thus $I_{t_{\max}}$ can be approximated by a local square integrable martingale with the following predictable variation process, see Section 3.2 in Andersen et al. (1985)

$$\begin{aligned} \langle I_{t_{\max}} \rangle \approx & \frac{n}{CIF_{1\theta}^4(t_{\max})} \times \\ & \left\{\int_0^{t_{\max}} \left[\frac{1-CIF_2(s)}{Y(s)} \int_s^{t_{\max}} dCIF_{1\theta}(u) - \frac{1}{Y(s)} \int_s^{t_{\max}} CIF_1(u) dCIF_{1\theta}(u)\right]^2 Y(s) d\Lambda_1(s) \right. \\ & \left. + \int_0^{t_{\max}} \left[\frac{CIF_1(s)}{Y(s)} \int_s^{t_{\max}} dCIF_{1\theta}(u) - \frac{1}{Y(s)} \int_s^{t_{\max}} CIF_1(u) dCIF_{1\theta}(u)\right]^2 Y(s) d\Lambda_2(s)\right\} \end{aligned}$$

Under the null hypothesis $H_0 : CIF_1(\cdot) \equiv CIF_{1\theta}(\cdot)$, $CIF_{1\theta}(\cdot)$ and $CIF_1(\cdot)$ can be used interchangeably i.e. calculation of quantities involving $CIF_1(\cdot)$ can be done by substituting $CIF_{1\theta}(\cdot)$ for $CIF_1(\cdot)$. Nevertheless I keep using $CIF_1(\cdot)$ as it simplifies notation. In addition, we can replace $CIF_2(\cdot)$, $\Lambda_2(\cdot)$ and $M_2(\cdot)$ in $I_{t_{\max}}$ and $\langle I_{t_{\max}} \rangle$ with their corresponding nonparametric estimates. Furthermore,

$$\int_s^{t_{\max}} dCIF_{1\theta}(u) = CIF_1(t_{\max}) - CIF_1(s)$$

and

$$\int_s^{t_{\max}} CIF_1(u) dCIF_{1\theta}(u) = \frac{1}{2}\left\{CIF_1^2(t_{\max}) - CIF_1^2(s)\right\}$$

Thus, we have the following estimators of $I_{t_{\max}}$ and $\langle I_{t_{\max}} \rangle$

$$\begin{aligned}\tilde{I}_{t_{\max},H_0} = \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \times & \left\{ \int_0^{t_{\max}} \left[\frac{1 - \widehat{CIF}_2(s)}{Y(s)} \left(CIF_1(t_{\max}) - CIF_1(s)\right) \right.\right. \\ & \left. - \frac{1}{2Y(s)} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right] dM_1(s) \\ & \left. + \int_0^{t_{\max}} \left[\frac{CIF_1(s) CIF_1(t_{\max})}{Y(s)} - \frac{1}{2Y(s)} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right] d\hat{M}_2(s) \right\}\end{aligned}$$

$$\begin{aligned}\left\langle \tilde{I}_{t_{\max},H_0} \right\rangle = \frac{n}{CIF_{1\theta}^4(t_{\max})} \times & \left\{ \int_0^{t_{\max}} \left[\left(1 - \widehat{CIF}_2(s)\right) \left(CIF_1(t_{\max}) - CIF_1(s)\right) \right.\right. \\ & \left. - \frac{1}{2} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right]^2 \frac{d\Lambda_1(s)}{Y(s)} \\ & \left. + \int_0^{t_{\max}} \left[CIF_1(s) CIF_1(t_{\max}) - \frac{1}{2} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right]^2 \frac{dA_2(s)}{Y(s)} \right\}\end{aligned}$$

Therefore, as an application of the martingale central limit theorem, see Theorem 5.1 in Kalbfleisch & Prentice (2002), under mild conditions, we have asymptotically

$$\frac{I_{t_{\max},H_0}}{\sqrt{\left\langle \tilde{I}_{t_{\max},H_0} \right\rangle}} = \frac{\sqrt{n}\left\{C_{t_{\max}} - \frac{1}{2}\right\}}{\sqrt{\left\langle \tilde{I}_{t_{\max},H_0} \right\rangle}} \sim \mathcal{N}(0,1) \; Q.E.D.$$

Alternatively, following Lin (1997) and Bajorunaite & Klein (2007) one can use directly the martingale representation of $I_{t_{\max}}$ to derive p-value based on resampling by replacing $dM_j(\cdot)$ with $\sum_{i=1}^n G_{ji} dN_{ji}(\cdot)$, $j = 1, 2$, where G_{ji}, $j = 1, 2$; $i = 1, ..., n$ are independent standard normal variables. This means

$\tilde{I}_{t_{\max},H_0}$ follows the same distribution as

$$\begin{aligned}\bar{I}_{t_{\max},H_0} =& \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \\ & \times \sum_{i=1}^{n} \left\{ \int_0^{t_{\max}} \left[\frac{1-\widehat{CIF}_2(s)}{Y(s)} \left(CIF_1(t_{\max}) - CIF_1(s)\right) \right.\right. \\ & \left. - \frac{1}{2Y(s)} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right] G_{1i} dN_{1i}(s) \\ & \left. + \int_0^{t_{\max}} \left[\frac{CIF_1(s) CIF_1(t_{\max})}{Y(s)} - \frac{1}{2Y(s)} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right] G_{2i} dN_{2i}(s) \right\}\end{aligned}$$

For a sample with observed event- or right-censoring times t_i, $i = 1, \ldots, n$, let I_1 and I_2 be, respectively, the index sets of subjects observed to have event type 1 and 2. If there are no two subjects failing from the same event type at the same time, $\tilde{I}_{t_{\max},H_0}$, $\bar{I}_{t_{\max},H_0}$ and $\left\langle \tilde{I}_{t_{\max},H_0} \right\rangle$ can be re-expressed as

$$\begin{aligned}\tilde{I}_{t_{\max},H_0} =& \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \times \\ & \sum_{i \in I_1} \left\{ \left[1 - \widehat{CIF}_2(t_i)\right] \left[CIF_1(t_{\max}) - CIF_1(t_i)\right] - \frac{1}{2} \left[CIF_1^2(t_{\max}) - CIF_1^2(t_i)\right] \right\} Y(t_i)^{-1} \\ & - \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \times \\ & \sum_{i=1}^{n} \int_0^{t_i} \left\{ \left[1 - \widehat{CIF}_2(s)\right] \left[CIF_1(t_{\max}) - CIF_1(s)\right] - \frac{1}{2} \left[CIF_1^2(t_{\max}) - CIF_1^2(s)\right] \right\} \frac{d\Lambda_1(s)}{Y(s)}\end{aligned}$$

$$\begin{aligned}\bar{I}_{t_{\max},H_0} =& \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \times \\ & \sum_{i \in I_1} \left\{ \left[1 - \widehat{CIF}_2(t_i)\right] \left[CIF_1(t_{\max}) - CIF_1(t_i)\right] - \frac{1}{2} \left[CIF_1^2(t_{\max}) - CIF_1^2(t_i)\right] \right\} \frac{G_{1i}}{Y(t_i)} \\ & + \frac{\sqrt{n}}{CIF_{1\theta}^2(t_{\max})} \times \\ & \sum_{i \in I_2} \left\{ CIF_1(t_i) CIF_1(t_{\max}) - \frac{1}{2} \left[CIF_1^2(t_{\max}) - CIF_1^2(t_i)\right] \right\} \frac{G_{2i}}{Y(t_i)}\end{aligned}$$

$$\begin{aligned}\left\langle \tilde{I}_{t_{\max},H_0} \right\rangle =& \frac{n}{CIF_{1\theta}^4(t_{\max})} \times \int_0^{t_{\max}} \left[\left(1 - \widehat{CIF}_2(s)\right) \left(CIF_1(t_{\max}) - CIF_1(s)\right) \right. \\ & \left. - \frac{1}{2} \left(CIF_1^2(t_{\max}) - CIF_1^2(s)\right) \right]^2 \frac{d\Lambda_1(s)}{Y(s)} \\ & + \frac{n}{CIF_{1\theta}^4(t_{\max})} \times \sum_{i \in I_2} \left[CIF_1(t_i) CIF_1(t_{\max}) - \frac{1}{2} \left(CIF_1^2(t_{\max}) - CIF_1^2(t_i)\right) \right]^2 Y(t_i)^{-2}\end{aligned}$$

B.2 Asymptotic $\bar{\chi}^2$-distributions

This provides a proof that extends the exact distributional assumptions and results of Shapiro (2003) to the case where both assumptions and results are asymptotic, which is necessary for the discussion in Subsection 5.3.2. In detail, Shapiro (2003) considered the case of a K-dimensional real-valued random vector X following an exact multivariate normal distribution $\mathcal{N}(0, V)$, where V is a nonsingular covariance matrix. In the notation of Subsection 5.3.1 X was $\mathcal{D} - \mathcal{D}_0$. Then for a closed and convex cone $\mathcal{C} \subset \mathbb{R}^k$, let $Z(\mathcal{C}, V) = \max_{w \in \mathcal{C}} \frac{(\mathcal{D}-\mathcal{D}_0)^T w}{(w^T V w)^{1/2}}$ then the random variate $[Z(\mathcal{C}, V)]^2$, according to Shapiro (2003) can also be expressed as

$$X^T V^{-1} X - \inf_{w \in \mathcal{C}} (X - w)^T V^{-1} (X - w)$$

which has an exact $\bar{\chi}^2$-distribution.

In what follows I shall prove the aforementioned asymptotic extension of this result as stated in the following lemma.

Lemma

If $\hat{X} \in \mathbb{R}^K$ is a statistic based on a sample of size n, e.g. in Subsection 5.3.2 $\hat{X} = \left(\hat{\mathcal{D}}(\tau) - \mathcal{D}_0(\tau)\right)$, following an asymptotic multivariate normal distribution $\mathcal{N}(0, V)$, where the nonsingular covariance matrix V can be consistently estimated by $n\hat{V}$. In formulas, these are

$$\sqrt{n}\hat{X} \xrightarrow{D} \mathcal{N}(0, V) \text{ and } n\hat{V} \xrightarrow{P} V$$

where $\xrightarrow{D}$ and $\xrightarrow{P}$ denote convergence in distribution and convergence in probability, respectively. Moreover all convergence is in sample size n and with respect to the Euclidean norm in $\mathbb{R}^K$ and some matrix norm in $\mathcal{V}^{K \times K}$, the vector space of $K \times K$ real symmetric matrices. It is also further assumed that for all n, $\hat{V}$ is symmetric and positive definite with probability one. Then

$$\hat{X}^T \hat{V}^{-1} \hat{X} - \inf_{w \in \mathcal{C}} (\hat{X} - w)^T \hat{V}^{-1} (\hat{X} - w) \xrightarrow{D} \bar{\chi}^2$$

Proof

First assume that the function $f : (y, V) \mapsto y^T V^{-1} y - \inf_{w \in \mathcal{C}} (y - w)^T V^{-1} (y - w)$ is continuous in (y, V) with respect to a suitably chosen norm defined on $\mathbb{R}^K \times \mathcal{V}^{K \times K}$. Then according to Theorem

2.7 in Van der Vaart (2000), $\sqrt{n}\hat{X} \xrightarrow{D} \mathcal{N}(0,V)$ and $n\hat{V} \xrightarrow{P} V$ imply $\left(\sqrt{n}\hat{X}, n\hat{V}\right) \xrightarrow{D} (X,V)$, where $X \sim \mathcal{N}(0,V)$. Thus, according to the continuous mapping theorem, we have

$$\left(\sqrt{n}\hat{X}\right)^T \frac{\hat{V}^{-1}}{n} \left(\sqrt{n}\hat{X}\right) - \inf_{w\in\mathcal{C}} (\sqrt{n}\hat{X}-w)^T \frac{\hat{V}^{-1}}{n} (\sqrt{n}\hat{X}-w) \xrightarrow{D} X^T V^{-1} X - \inf_{w\in\mathcal{C}} (X-w)^T V^{-1} (X-w)$$

It can easily be seen that for all n, the left hand side of the above equation equals

$$\hat{X}^T \hat{V}^{-1} \hat{X} - \inf_{w\in\mathcal{C}} (\hat{X}-w)^T \hat{V}^{-1} (\hat{X}-w)$$

where $\inf_{w\in\mathcal{C}} (\sqrt{n}\hat{X}-w)^T \frac{\hat{V}^{-1}}{n} (\sqrt{n}\hat{X}-w)$ equals $\inf_{w\in\mathcal{C}} (\hat{X}-w)^T \hat{V}^{-1} (\hat{X}-w)$ due to the definition of the cone $\mathcal{C}$. As a result we have

$$\hat{X}^T \hat{V}^{-1} \hat{X} - \inf_{w\in\mathcal{C}} (\hat{X}-w)^T \hat{V}^{-1} (\hat{X}-w) \xrightarrow{D} \bar{\chi}^2$$

It remains to verify the continuity of $f(\cdot,\cdot)$ as assumed earlier.

Let us continue by first specifying the appropriate norms for each variable. For y in $\mathbb{R}^K$ I use the normal Euclidean norm $\|\cdot\|_2$, and for all $V \in \mathcal{V}^{K\times K}$, let $\|V\|_\infty = \max_{1\le i,j\le K} |V_{ij}|$. This matrix norm is chosen simply for ease of discussion and continuity of $f(\cdot,\cdot)$ once proven should hold regardless of the choice of matrix norm. Then the corresponding norm for $(y,V,w) \in \mathbb{R}^K \times \mathcal{V}^{K\times K} \times \mathbb{R}^K$ is $\|(y,V,w)\| = \|y\|_2 + \|V\|_\infty + \|w\|_2$. Moreover, I shall drop all subscripts when referring to a norm for simplicity and any topological statements involving a variable should be implicitly understood as being considered with respect to the topology generated by the norm defined for that variable. Then it is simple to prove that V^{-1} is a continuous function of V if V is nonsingular and that $y^T V y$ and $(y-w)^T V (y-w)$ are continuous in (y,V) and (y,V,w), respectively. Thus we are only left with showing that $\inf_{w\in\mathcal{C}} (y-w)^T V (y-w)$ is continuous in (y,V).

For this, note that $\inf_{w\in\mathcal{C}} (y-w)^T V (y-w)$ is the projection of y onto a closed and convex set $\mathcal{C}$ with respect to the strictly convex norm $\|y-w;V\| = (y-w)^T V (y-w)$ (as long as V is positive definite). Thus, according to Section 8.1 in Boyd & Vandenberghe (2004) for any positive definite matrix $V \in \mathcal{V}^{K\times K}$ and any $y \in \mathbb{R}^K$ there exists exactly one $w_{y,V}$ in $\mathcal{C}$ such that $(y-w_{y,V})^T V (y-w_{y,V}) = \inf_{w\in\mathcal{C}} (y-w)^T V (y-w)$ i.e. the minimizer is uniquely attained. To this end, for simplicity let $g(y,V,w) = (y-w)^T V (y-w)$ and $h(y,V) = \min_{w\in\mathcal{C}} g(y,V,w)$.

For any $(y_0,V_0) \in \mathbb{R}^K \times \mathcal{V}^{K\times K}$ and any sequence $\{y_n,V_n\}$ in $\mathbb{R}^K \times \mathcal{V}^{K\times K}$ converging to (y_0,V_0); for each n let $w_n = \operatorname{argmin}_{w\in\mathcal{C}} g(y_n,V_n,w)$ and $w_0 = \operatorname{argmin}_{w\in\mathcal{C}} g(y_0,V_0,w)$. If $\mathcal{C}$ is compact, we have that $h(y_n,V_n)$ converging to $h(y_0,V_0)$ implies the continuity of $h(\cdot,\cdot)$. However, a cone cannot be compact. The main idea of what follows is to show that the sequence $\{w_n\}$ is bounded.

Thus there exists a closed, bounded and hence compact $\mathcal{C}_0 \subset \mathcal{C}$ covering w_0 and $\{w_n\}$. In such case, $\min_{w\in\mathcal{C}} g(y_n, V_n, w) = \min_{w\in\mathcal{C}_0} g(y_n, V_n, w)$ for all n, and $\min_{w\in\mathcal{C}} g(y_0, V_0, w) = \min_{w\in\mathcal{C}_0} g(y_0, V_0, w)$. Thus one can argue for the continuity of $h(\cdot,\cdot)$ as in the case when $\mathcal{C}$ is compact.

Now assume the contrary that $\{w_n\}$ is unbounded. Thus there is a subsequence $\{w_{k_n}\}$ such that $\lim_{n\to\infty} \|w_{k_n}\| = \infty$. This means for large enough n, $\|w_{k_n}\| > 0$, and for such n the sequence $\left\{w_{k_n} \|w_{k_n}\|^{-1}\right\} \subseteq \mathcal{C}$ (as $\mathcal{C}$ is a cone) is bounded and thus has its own subsequence that converges to $\tilde{w} \neq 0$ in $\mathcal{C}$ (due to $\mathcal{C}$'s closedness). With a slight abuse of notation, let $\left\{w_{k_n} \|w_{k_n}\|^{-1}\right\}$ also denote this subsequence.

For all n, due to the definition of w_n we have $g(y_{k_n}, V_{k_n}, w_{k_n}) \leq g(y_{k_n}, V_{k_n}, w_0)$. Thus,

$$\begin{aligned}
&\forall n, \frac{g(y_{k_n}, V_{k_n}, w_{k_n})}{\|w_{k_n}\|^2} \leq \frac{g(y_{k_n}, V_{k_n}, w_0)}{\|w_{k_n}\|^2} \\
&\Rightarrow \lim_{n\to\infty} \frac{g(y_{k_n}, V_{k_n}, w_{k_n})}{\|w_{k_n}\|^2} \leq \lim_{n\to\infty} \frac{g(y_{k_n}, V_{k_n}, w_0)}{\|w_{k_n}\|^2} \\
&\Leftrightarrow \lim_{n\to\infty} g(0, V_0, \tilde{w}) \leq 0
\end{aligned}$$

This means 0 is not the only minimizer of $g(0, V_0, w)$ in $\mathcal{C}$ which is a contradiction $Q.E.D.$