The Analysis of Competing Risks Data

with a Focus on Estimation of Cause-Specific and Subdistribution Hazard Ratios from a Mixture Model

Bernhard Haller

Dissertation an der
Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

05. Februar 2014

Erstgutachter: Prof. Dr. Kurt Ulm
Zweitgutachter: Prof. Dr. Thomas Augustin
Drittgutachter: Prof. Dr. Stefan Wagenpfeil
Tag der Disputation: 20.05.2014

Danksagung

Ich möchte mich herzlich bei Prof. Dr. Kurt Ulm vom Institut für Medizinische Statistik und Epidemiologie (IMSE) der Technischen Universität München und bei Prof. Dr. Thomas Augustin vom Institut für Statistik der Ludwig-Maximilians-Universität München für die Ermöglichung dieser Arbeit und die umfangreiche Unterstützung bedanken. Darüber hinaus geht mein Dank an meine Kollegen am IMSE und an die Mitglieder der Arbeitsgruppe Method(olog)ische Grundlagen der Statistik und ihre Anwendungen für die ertragreichen Diskussionen, die geholfen haben, diese Arbeit fertigzustellen und zu verbessern. Bei Prof. Dr. Georg Schmidt möchte ich mich für die Bereitstellung der verwendeten Daten erkenntlich zeigen.

Ein besonderer Dank gilt meiner Familie für ihre Geduld und ihre Unterstützung während der Entstehung dieser Arbeit, ohne die ein erfolgreicher Abschluss nicht möglich gewesen wäre.

Summary

Treatment efficacy in clinical trials is often assessed by time from treatment initiation to occurrence of a certain critical or beneficial event. In most cases the event of interest cannot be observed for all patients, as patients are only followed for a limited time or contact to patients is lost during their follow-up time. Therefore, certain methods were developed in the framework of the so called time-to-event or survival analysis, in order to obtain valid and consistent estimates in the presence of these "censored observations", using all available information. In classical event time analysis only one endpoint exists, as the death of a patient. As patients can die from different causes, in some clinical trials time to one out of two or more mutually exclusive types of event may be of interest. In many oncological studies, for example, time to cancer-specific death is considered as primary endpoint with deaths from other causes acting as so called competing risks. Different methods for data analysis in the competing risks framework were developed in recent years, which either focus on modelling the cause-specific or the subdistribution hazard rate or split the joint distribution of event times and event types into quantities, that can be estimated from observable data. In this work the analysis of event time data in the presence of competing risks is described, including the presentation and discussion of different regression approaches. A major topic of this work is the estimation of cause-specific and subdistribution hazard rates from a mixture model and a new approach using penalized B-splines (P-splines) for estimation of conditional hazard rates in a mixture model is proposed. In order to evaluate the behaviour of the new approach, a simulation study was conducted, using simulation techniques for competing risks data, which are described in detail in this work. The presented regression models were applied to data from a clinical cohort study investigating a risk stratification for cardiac mortality in patients, that survived a myocardial infarction. Finally, the use of the presented methods for event time analysis in the presence of competing risks and results obtained from the simulation study and the data analysis are discussed.

Zusammenfassung

Zur Beurteilung der Wirksamkeit von Behandlungen in klinischen Studien wird häufig die Zeit vom Beginn einer Behandlung bis zum Eintreten eines bestimmten kritischen oder erwünschten Ereignisses als Zielgröße verwendet. Da in vielen Fällen das entsprechende Ereignis nicht bei allen Patienten beobachtet werden kann, da z.B. Patienten nur für einen gewissen Zeitraum nachverfolgt werden können oder der Patientenkontakt in der Nachbeobachtungszeit abbricht, wurden im Rahmen der so genannten Ereigniszeit- bzw. Überlebenszeitanalyse Verfahren entwickelt, die bei Vorliegen dieser „zensierten Beobachtungen“ konsistente Schätzer liefern und dabei die gesamte verfügbare Information verwenden. In der klassischen Ereigniszeitanalyse existiert nur ein möglicher Endpunkt, wie der Tod eines Patienten. Da Patienten jedoch an verschiedenen Ursachen versterben können, ist in manchen klinischen Studien die Zeit bis zu einem von zwei oder mehreren sich gegenseitig ausschließenden Ereignistypen von Interesse. So fungiert z.B. in vielen onkologischen Studien die Zeit bis zum tumor-bedingten Tod als primärer Endpunkt, wobei andere Todesursachen sogenannte konkurrierende Risiken („Competing Risks“) darstellen. In den letzten Jahren wurden mehrere Verfahren zur Datenanalyse bei Vorliegen konkurrierender Risiken entwickelt, bei denen entweder die ereignis-spezifische oder die Subdistribution-Hazardrate modelliert wird, oder bei denen die gemeinsame Verteilung von Ereigniszeiten und Ereignistypen als Produkt von Größen abgebildet wird, die aus den beobachtbaren Daten geschätzt werden können. In dieser Arbeit werden Methoden zur Analyse von Competing-Risks-Daten, einschließlich verschiedener Regressionsansätze, vorgestellt. Besonderes Augenmerk liegt auf der Schätzung der ereignis-spezifischen und Subdistribution-Hazardraten aus einem sogenannten Mixture Model. Diesbezüglich wird auch ein neuer Ansatz zur Schätzung der konditionalen Hazardraten in einem Mixture Model unter Verwendung penalisierter B-Spline-Funktionen (P-Splines) vorgestellt. Um die Eigenschaften des neuen Ansatzes zu untersuchen, wurde eine Simulationsstudie unter Einsatz verschiedener Simulationsstrategien für Competing-Risks-Daten, die in dieser Arbeit im Detail beschrieben werden, durchgeführt. Die Regressionsmodelle wurden auf Daten einer klinischen Kohortenstudie zur Evaluation einer Risikostratifizierung für Patienten, die einen Myokardinfarkt überlebt haben, angewandt. Abschließend werden die vorgestellten Methoden zur Analyse von Ereigniszeitdaten bei Vorliegen konkurrierender Risiken sowie die Ergebnisse der Simulationsstudie und der Datenanalyse diskutiert.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In clinical research time from a certain starting point, as diagnosis of a disease or treatment initiation, to occurrence of a critical or beneficial event is often used to assess efficacy of a certain therapy or the role of potential predictive or prognostic factors. In most cases event times cannot be observed for all individuals due to a limited follow-up time or as patients are lost to follow-up. In order to obtain unbiased estimates for event time distributions or for covariate effects on event times, using the whole information available for the study cohort under investigation, methods that can deal with these "censored observations" were developed in the framework of the so called time-to-event or survival analysis. In standard survival analysis each subject can fail from one possible endpoint, so the information obtained for each subject is either his event time or a last observation time the subject was known to be event-free. Various textbooks on event time analysis are available, as e.g. Marubini and Valsecchi (1995), Kalbfleisch and Prentice (2002) or Klein and Moeschberger (2003).

Since certain treatments or risk factors only have an effect on one specific endpoint, the main interest may be on time to a certain type of event. In oncological trials, intended to compare different treatment strategies, the primary endpoint often is "time to tumour-related death", since it is expected that "time to other causes of death" is not affected or affected in a different way by the therapies. So each subject can fail from one out of two or more possible types of event and times to different event types can either be of same importance or time to one certain type of event may be of major interest. In presence of these so called "competing risks" the application of standard methods for event time analysis, which were developed for common survival analysis with one possible type of event, and ignoring the competing events may give erroneous results. As in the competing risks framework different event types are considered to be mutually exclusive, only the time to the first of these events can be observed. So the joint distribution of event times for different types of event cannot be estimated from the observed data due to non-identifiability problems.

Although the issue of competing risks in survival analysis was recognized as early as 1760 by Daniel Bernoulli, who tried to account for patients deaths from other causes in an investigation of the efficacy of pock immunization, which was published by Bernoulli in 1766, translated by Bradley (1971), and revisited by Dietz and Heesterbeek (2002), wrong or inadequate methods are still in use for analysis and presentation of competing risks data in clinical research and consequently in medical literature. E.g. the popular method proposed

by Kaplan and Meier (1958), which is commonly used for estimation of survival probabilities in standard event time analysis, leads to biased estimates of event probabilities when competing risks are treated as censored observations. Nevertheless, application of that procedure can be found in many medical publications with present competing risks. A systematic review of medical articles with probable competing risks endpoints in medical journals with high impact factors, which was performed by Koller et al (2012), revealed, that competing risks were not correctly analysed or that methods and results of the analyses were not adequately presented in many of the investigated articles, although a variety of publications describing methods for analysis of competing risks data is available in the statistical literature. Standard methods for the analysis of competing risks data are described and discussed in various introductory articles in statistical journals (Putter et al, 2007; Klein, 2010; Bakoyannis and Touloumi, 2011), in textbooks on competing risks (Pintilie, 2006; Beyersmann et al, 2012) or in chapters of textbooks on survival analysis (Marubini and Valsecchi, 1995; Kalbfleisch and Prentice, 2002). In recent years the competing risks problem was recognized in some medical disciplines, leading to articles dealing with issues to be considered in the presence of competing risks, that were published in medical journals, mainly in journals with a focus on cancer research (e.g. Satagopan et al, 2004; Kim, 2006; Dignam and Kocherginsky, 2008; Dignam et al, 2012), but also in journals for geriatrics (Berry et al, 2010) or urology (Roobol and Heinsdijk, 2011), including coverage of software applications for analysis of competing risks data (Scrucca et al, 2007).

Generally, methods considering the presence of multiple event types have to be conducted for the analysis of competing risks data. Different approaches for analysis of event time data with multiple, mutually exclusive types of event were proposed in the past. Early analysis of competing risks data focussed on estimation of the joint distribution of event times to different types of event, as e.g. the models for exponentially distributed event times with two types of failure, which were introduced by Cox (1959). Tsiatis (1975) demonstrated that this joint distribution cannot be estimated in a competing risks setting with mutually exclusive types of event without making strong, unverifiable assumptions about the dependence structure between times to different event types, and that for observed marginal event time distributions different joint distributions can be found. Prentice et al (1978) proposed to focus on the so called cause-specific hazard rate, the adaptation of the common hazard rate for the competing risks setting, which can be estimated from observable data. They introduced a Cox-type regression model to assess covariate effects on the cause-specific hazard rates. When cause-specific hazards are modelled, the estimated probability for an event of interest up to a given time, represented by the so called cumulative incidence function in the competing risks setting, depends on the cause-specific hazard rates for all possible types of event, as these have an influence on the number of patients at risk (see e.g. Putter et al, 2007). In 1985 Larson and Dinse proposed to represent the non-estimable joint distribution of event times and types of event by the product of the marginal event type distribution, assessed e.g. via a multinomial logistic regression model, and the conditional event time distributions given the type of event, using parametric survival models. In their original article a piecewise exponential model was proposed to assess the conditional event time distributions. A common likelihood can be denoted for the model and the parameters can be estimated from observable data using numerical approaches for maximum likelihood estimation. In order to find a "hazard-like" quantity in the presence of competing risks, that is directly linked to the cumulative incidence function

as known from standard survival analysis, Gray (1988) introduced the so called subdistribution hazard. For the subdistribution hazard an adapted risk set is considered, keeping individuals that failed from a competing event in the risk set for future timepoints. The approach was extended to a regression model by Fine and Gray (1999) allowing to assess the influence of covariates on the subdistribution hazard. In this approach, the coefficients obtained e.g. from a Cox-type regression model are monotonously linked to the cumulative incidence function, so covariate effects derived from a subdistribution hazards regression model can be translated directly to effects on event probabilities, given the model assumptions hold. Andersen et al (2003) introduced a method for estimation of covariate effects on a quantity of interest in survival models using pseudo-values. The method was adjusted later for the competing risks framework by using the cumulative incidence function as measure of interest (Klein and Andersen, 2005). The generalized estimating equation approach (GEE) by Liang and Zeger (1986) is used to estimate the influence of covariates on the cumulative incidence function and to give robust standard errors leading to valid p values and confidence intervals. In 2010, Nicolaie et al proposed another way of factorizing the joint distribution of event times and types of event by expressing the joint distribution as product of the marginal event time distribution and the conditional distribution of event types given the time of event. The so called vertical modelling approach, consisting e.g. of a parametric survival model to assess the marginal event time distribution and a multinomial logistic regression model for the conditional event type distribution, using time and covariates of interest as independent variables, provides estimates for relative hazard rates, which represent the pattern of events over the course of time.

Most competing risks analyses presented in the medical literature are performed hazard-based. Often Cox-type regression models for the cause-specific or the subdistribution hazard rate are considered, assuming proportional cause-specific or subdistribution hazards, respectively. It was shown, that the proportionality assumption generally does not hold for both quantities in the presence of competing risks (Beyersmann and Schumacher, 2007; Grambauer et al, 2010). In a recent article by Lau et al (2011) the estimation of cause-specific and subdistribution hazard rates and consequently hazard ratios from a mixture model, assuming the conditional event times to follow flexible parametric event time distributions, as the three-parameter generalized gamma distribution, was presented and applied to a dataset. The method is intended to provide estimates for cause-specific and subdistribution hazard rates from one common model, to detect time-dependencies of cause-specific and subdistribution hazard rates and hazard ratios, and to allow for estimation of average hazard ratios. The procedure is described in this work and a new approach using penalized B-splines (P-splines) for estimation of conditional hazard rates is proposed, as the approach considering the generalized gamma distribution was found to be numerically unstable. In order to evaluate the new approach and to compare it to the models, that were proposed by Lau et al, a simulation study was conducted for different scenarios with prespecified time-constant or time-dependent cause-specific and subdistribution hazard rates and hazard ratios using strategies for simulation of competing risks data (Beyersmann et al, 2009).

In this work the competing risks framework is described and problems occurring for analysis of event time data in the presence of multiple, mutually exclusive types of event are presented. In Section 2 an overview over standard event time analysis with one possible endpoint is given and relevant quantities and concepts are introduced. The competing risks

setting is described in Section 3, including discussions on different views on the competing risks situation and on the non-identifiability problem, and presentation of relevant quantities used for description of competing risks data. In Section 4 regression models for the competing risks setting, which were proposed in the literature, are described and compared regarding model assumptions, applicability, and interpretation of obtained results. Various extensions of the basic models are mentioned and literature for further reading is given. A special focus lies on the derivation of estimates for cause-specific and subdistribution hazard rates and hazard ratios from a mixture model. In Section 5 the new mixture model approach using penalized B-spline basis functions (P-splines) for estimation of conditional hazard rates is introduced. Different methods for simulation of competing risks data following predefined cause-specific or subdistribution hazards are presented and discussed in Section 6. In Section 7 a simulation study, performed to investigate the properties of different mixture models for estimation of cause-specific and subdistribution hazard rates and hazard ratios, is described. The focus lies on a comparison of the newly proposed mixture model approach, using P-spline functions for estimation of conditional hazard rates, and parametric mixture models. Different scenarios using predefined cause-specific or subdistribution hazard rates and different censoring distributions were considered. The methods under investigation were compared regarding numerical stability and ability to detect the true underlying hazard rates and hazard ratios. In Section 8 application of the presented methods for competing risks regression and of the newly proposed mixture model approach to data from a clinical cohort study, investigating risk stratification for cardiac death after myocardial infarction, is described. Finally, the presented methods as well as the findings from the simulation study and the data analyses are discussed in Section 9. All analyses and the simulation study were performed using the statistical software R (R Development Core Team, 2011). Sketches of R-codes used for data analysis and simulation as well as further results of the simulation study are presented in the Appendix.
Description of the different competing risks regression models and their application to the clinical cohort study presented in Sections 4 and 8 were published in the journal *Lifetime Data Analysis* (Haller et al, 2013). Presentation and validation of the Binomial Algorithm for generation of competing risks data following a predefined subdistribution hazard ratio described in Sections 6.3.3 and 6.3.4 were published in the *Journal for Statistical Computation and Simulation* (Haller and Ulm, 2013). The proposed mixture model approach using penalized B-splines for estimation of conditional hazard rates, intended to derive cause-specific and subdistribution hazard rates, which is introduced in Section 5, the results of the simulation study for Scenarios II to IV with a moderate amount of censored observations presented in Section 7, and application of the P-spline approach to the clinical cohort study and according results (Section 8.3), are described in a manuscript that was submitted for publication and was under review at the time this work was finalized.

Chapter 2

Analysis of time-to-event data

The main topic of this work, the analysis of competing risks data, is a special case of event time or survival analysis. Time-to-event data are often considered in the medical context or in quality assurance, investigating e.g. the time from therapy onset to time of death or time of full recovery, or evaluating the time until a machine breaks down and has to be replaced. There are also applications of event time data analysis methods in social or financial sciences, investigating e.g. the time an individual is unemployed or the time a company stays on the market.
When time-to-event data are considered, often the time to the event of interest cannot be observed for all individuals or subjects, as either subjects have not experienced an event at the end of the study (administrative censoring), subjects drop-out early from the study, or are lost to follow-up. These observations are called "censored observations" in the time-to-event framework. Special methods that are able to deal with these censored observations were introduced in the context of event time or survival analysis, in order to obtain unbiased estimates for relevant quantities without losing information. While methods for consideration of dependent censoring times, i.e. censoring times that depend on covariate values, were introduced in the literature (see e.g. Robins and Finkelstein, 2000), only non-informative censoring is assumed in this work, i.e. the censoring times are assumed to be stochastically independent of observed and unobserved covariates and of the event times. The most important measures and issues relevant for the analysis of time-to-event data are summarized in this section. Further details on the analysis of event time data can be found in various textbooks (see e.g. Marubini and Valsecchi, 1995; Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2003).

2.1 Observed data

Observed time-to-event data can be represented by a pair of variables, the observed time T^* and a status variable D. For each individual only the minimum of the true event time, denoted by the random variable T, and the potential censoring time C can be observed

$$T^* = \min\{T, C\}.$$

The status variable D indicates, whether the observed time T^* is a real event time or a censoring time

$$D = I(T < C),$$

with $I(\cdot)$ being the indicator function returning a value of one, if the given expression is true, and zero else. So for each individual $i=\{1,\ldots,n\}$ a couple (t_i, d_i) is observed.

2.2 Important measures

The random variable for the event time T is strictly positive and its distribution is defined by the density function $f(t)$ or the cumulative density function $F(t)$ with

$$F(t) = P(T \leq t) = \int_0^t f(s)\mathrm{d}s. \tag{2.1}$$

In the context of event time analyses, data are often presented by the so called survivor function

$$S(t) = P(T > t) = 1 - F(t), \tag{2.2}$$

denoting the probability that an individual is event-free up to a given time t.
The survivor function can be estimated by the popular Kaplan-Meier method (Kaplan and Meier, 1958). Denoting the vector of distinct observed event times as $\tilde{\boldsymbol{t}} = (\tilde{t}_1, \ldots, \tilde{t}_N)$, with N being the number of distinct timepoints with an observed event, and $d_{\tilde{t}_i}$ and $R_{\tilde{t}_i}$ as the number of events observed at $\tilde{t}_i$ and the number of individuals at risk at $\tilde{t}_i$, an estimate for the survivor function at a given time t can be obtained as

$$\hat{S}(t) = \prod_{i:\,\tilde{t}_i \leq t} \left(1 - \frac{d_{\tilde{t}_i}}{R_{\tilde{t}_i}}\right). \tag{2.3}$$

The Kaplan-Meier estimator returns a step function with jumps at observed event times and constant estimates of the survivor function for timepoints between observed event times.
The hazard rate $\lambda(t)$ indicates the probability for an event in an infinitesimal small time interval $t+\Delta t$, given the subject did not fail before t,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.4}$$

For a timepoint with an observed event $\tilde{t}_i$ the hazard rate can be estimated as the number of observed events at $\tilde{t}_i$, denoted by $d_{\tilde{t}_i}$, divided by the number of subjects under risk at the time of interest

$$\hat{\lambda}(\tilde{t}_i) = \frac{d_{\tilde{t}_i}}{R_{\tilde{t}_i}}. \tag{2.5}$$

For timepoints without an observed event the non-parametric estimator for the hazard rate returns zero. Therefore, presentation of the hazard function is mostly conducted showing an estimate for the cumulative hazard rate $\Lambda(t)$, which is defined as the integral over the hazard function from zero to the time of interest t,

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s. \tag{2.6}$$

The cumulative hazard rate can be estimated using the Nelson-Aalen estimator (Nelson, 1969), summing up the quotients of observed events and the numbers of subjects at risk, as described above, for all observed event times up to the time of interest t, leading to a monotonically increasing step function,

$$\hat{\Lambda}(t) = \sum_{i:\,\tilde{t}_i \leq t} \frac{d_{\tilde{t}_i}}{R_{\tilde{t}_i}}. \tag{2.7}$$

The density function $f(t)$, the survivor function $S(t)$ and the hazard function $\lambda(t)$ are directly related via

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{2.8}$$

and

$$S(t) = 1 - F(t) = \exp\big(-\Lambda(t)\big). \tag{2.9}$$

2.3 Regression models for time-to-event data

In order to investigate the influence of covariates on the event times, regression models that can be applied in the presence of censored observations were developed. Commonly used regression models for time-to-event data with one possible endpoint are summarized in this section.

2.3.1 Cox regression

The most popular regression model for event time data is the proportional hazards model introduced by Cox (1972). It is assumed that hazard ratios are constant over time and that each of the P covariates under consideration has a linear effect on the logarithm of the hazard rate, given the other covariates. The Cox regression model can be written as

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\, \exp(\boldsymbol{\beta}^\top \mathbf{x}), \tag{2.10}$$

with the unspecified baseline hazard rate $\lambda_0(t)$ for a (possibly fictitious) individual with a covariate vector of zeros, the P-dimensional vector of covariates $\mathbf{x}$ and the vector of regression coefficients $\boldsymbol{\beta}$. The hazard ratio between two individuals i and j can be computed as

$$\frac{\lambda(t|\mathbf{x_j})}{\lambda(t|\mathbf{x_i})} = \frac{\lambda_0(t)\, \exp(\boldsymbol{\beta}^\top \mathbf{x_j})}{\lambda_0(t)\, \exp(\boldsymbol{\beta}^\top \mathbf{x_i})} = \exp\big(\boldsymbol{\beta}^\top (\mathbf{x_j} - \mathbf{x_i})\big). \tag{2.11}$$

Consequently, the regression coefficient for the p^{th} covariate β_p can be interpreted as the logarithm of the hazard ratio between two individuals, differing in one unit of the covariate x_p, and having equal values for all other covariates x_q with $q \neq p$

$$\beta_p = \ln\left(\frac{\lambda(t|x_1, x_2, \ldots, x_{p-1}, x_p{+}1, x_{p+1}, \ldots, x_{P-1}, x_P)}{\lambda(t|x_1, x_2, \ldots, x_{p-1}, x_p, x_{p+1} \ldots, x_{P-1}, x_P)}\right). \tag{2.12}$$

Assuming N distinct ordered failure times $\tilde{\boldsymbol{t}}{=}(\tilde{t}_1, \ldots, \tilde{t}_N)$, an estimate for the vector of regression coefficients $\boldsymbol{\beta}$ is derived by numerical maximization of the partial likelihood

introduced by Cox, treating the unspecified baseline hazard $\lambda_0(t)$ as a nuisance parameter

$$PL = \prod_{i=1}^{N} \frac{\exp(\mathbf{x_{(i)}}^\top \boldsymbol{\beta})}{\sum\limits_{j \in R_{\tilde{t}_i}} \exp(\mathbf{x_j}^\top \boldsymbol{\beta})}, \tag{2.13}$$

where $\mathbf{x_{(i)}}$ is the vector of covariates of the individual that failed at time $\tilde{t}_i$, and $R_{\tilde{t}_i}$ represents the risk set at event time $\tilde{t}_i$, i.e. all subjects not having failed before $\tilde{t}_i$ and still under observation at $\tilde{t}_i$. Usually, a Newton-Raphson algorithm is performed to find the vector of regression coefficients, that maximizes the log-partial likelihood

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \ln\big(PL(\boldsymbol{\beta})\big) \right\}. \tag{2.14}$$

The variance of the estimated regression coefficients is provided by the inverse of the observed information matrix evaluated at the maximum partial likelihood estimate

$$Var(\hat{\beta}) = \mathbf{I}(\hat{\beta})^{-1}, \tag{2.15}$$

which can be obtained as the negative of the second derivative of the log-partial likelihood function shown in Equation 2.13.
The assumption of linear covariate effects on the log-hazard rate can be checked using deviance or martingale residuals (see e.g. Therneau and Grambsch, 2000). Time-dependence of regression coefficients can be assessed by Schoenfeld residuals (Schoenfeld, 1982). A more detailed description of the Cox regression model including statistical tests for regression coefficients as well as further extensions of the model, as e.g. stratification using different baseline hazard functions for predefined groups or frailty models allowing for random effects, can be found in various textbooks (Marubini and Valsecchi, 1995; Therneau and Grambsch, 2000; Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2003).

2.3.2 Parametric regression models

As an alternative to Cox regression, parametric survival models, assuming the event times to follow a prespecified distribution and using an adequate link between the linear predictor $\boldsymbol{\beta}^\top \mathbf{x}$ and the parameters of the event time distribution, can be used. Popular event time distributions are e.g. the exponential distribution, the Weibull distribution or the log-normal distribution. Event time distributions, which are used later in this work are presented in Section 2.4.
In a parametric survival model the likelihood, which has to be maximized in order to obtain parameter estimates, is

$$L = \prod_{i=1}^{n} \left(\big[f(t_i|\mathbf{x_i})\big]^{I(d_i=1)} \big[S(t_i|\mathbf{x_i})\big]^{I(d_i=0)} \right), \tag{2.16}$$

with d_i being the status variable described in Section 2.1, indicating whether an observed time t_i is a real failure time (d_i=1) or a last time a subject was known to be event-free

(d_i=0), $f(t)$ being the density function of the event time distribution, and $S(t)$ denoting the corresponding survivor function. Consequently, the log-likelihood function is

$$\begin{aligned} ll &= \sum_{i=1}^{n} \Big[I(d_i{=}1)\ \ln\big(f(t_i|\mathbf{x_i})\big) + I(d_i{=}0)\ \ln\big(S(t_i|\mathbf{x_i})\big)\Big] = \\ &= \sum_{i=1}^{n} \Big[I(d_i{=}1)\Big(\ln\big(\lambda(t_i|\mathbf{x_i})\big) + \ln\big(S(t_i|\mathbf{x_i})\big)\Big) + I(d_i{=}0)\ \ln\big(S(t_i|\mathbf{x_i})\big)\Big] = \\ &= \sum_{i=1}^{n} \Big[I(d_i{=}1)\ \ln\big(\lambda(t_i|\mathbf{x_i})\big) + \ln\big(S(t_i|\mathbf{x_i})\big)\Big]. \end{aligned} \tag{2.17}$$

Estimates for the parameters are derived by maximizing the likelihood or the log-likelihood function either analytically or numerically. The variance of the maximum likelihood estimates can be derived from the inverse of the observed Fisher information matrix as common for parametric regression models (see e.g. Fahrmeir and Tutz, 2001).
Very flexible parametric regression models, e.g. assuming the event times to follow a three-parameter generalized gamma distribution (Cox et al, 2007, see also Section 2.4.3), were introduced in recent years. Rosenberg (1995) or Royston and Parmar (2002) presented approaches for flexible estimation of hazard rates by inclusion of cubic spline functions (see also Section 5).

2.4 Event time distributions

In this section some event time distributions, which are commonly considered for parametric survival models and which are used later in this work for estimation of conditional event time distributions in competing risks mixture models, are summarized. An overview over common event time distributions including important quantities as density functions, survivor functions, and hazard functions can be found in Chapter 2 of the textbook by Klein and Moeschberger (2003). In contrary to the Cox regression model, the vector of regression coefficients $\mathbf{x}$ in a parametric regression model usually includes an intercept term, which is not mentioned explicitly in the following.

2.4.1 Exponential distribution

The exponential distribution is a one-parametric event time distribution implying a time-constant hazard rate $\lambda(t)$=λ. The density function of the exponential distribution is

$$f(t) = \lambda \exp(-\lambda t) \tag{2.18}$$

and the survivor function is

$$S(t) = \exp(-\lambda t). \tag{2.19}$$

In a regression model investigating the influence of covariates on the event times, the hazard rate λ is commonly modelled via

$$\lambda = \exp(\boldsymbol{\beta}^\top \mathbf{x}), \tag{2.20}$$

to ensure positivity of the estimated hazard rates.

2.4.2 Weibull distribution

The Weibull distribution is defined by the two parameters λ and α, and therefore it is more flexible than the exponential distribution. Different parametrizations of the Weibull distribution exist in the literature. One possible formulation of the density function, which is used throughout this work, is

$$f(t) = \lambda\alpha(\lambda t)^{\alpha-1} \exp\big(-(\lambda t)^{\alpha}\big), \tag{2.21}$$

so the exponential distribution is a special case of the Weibull distribution for α=1. The survivor function for that parametrization is

$$S(t) = \exp\big(-(\lambda t)^{\alpha}\big) \tag{2.22}$$

and the hazard function can be denoted as

$$\lambda(t) = \lambda\alpha(\lambda t)^{\alpha-1}. \tag{2.23}$$

For regression purposes the influence of the covariates on the parameter λ as described in Equation 2.20 or on both parameters λ and α is assessed.

2.4.3 Generalized gamma distribution

A flexible parametric survival model is the generalized gamma model, that was e.g. used by Cox et al (2007) for comparison of different treatment eras regarding time from diagnosis of AIDS to death in HIV positive patients. The density function of the generalized gamma distribution with three parameters can be denoted as

$$f(t) = \frac{|\nu|}{\tilde{\alpha}t\Gamma(\nu^{-2})} \Big(\nu^{-2}(\lambda t)^{\nu/\tilde{\alpha}}\Big)^{\nu^{-2}} \exp\Big(-\nu^{-2}(\lambda t)^{\nu/\tilde{\alpha}}\Big), \tag{2.24}$$

where $\Gamma(\cdot)$ is the gamma function. The corresponding survivor function for the generalized gamma distribution is

$$\begin{aligned} S(t) &= 1 - F_{\Gamma}\big(\nu^{-2}(\lambda t)^{\nu/\tilde{\alpha}}; \nu^{-2}\big) \qquad \text{for } \nu > 0, \\ S(t) &= F_{\Gamma}\big(\nu^{-2}(\lambda t)^{\nu/\tilde{\alpha}}; \nu^{-2}\big) \qquad \text{for } \nu < 0, \end{aligned} \tag{2.25}$$

with $F_{\Gamma}(t, y)$ being the cumulative density function of the two parameter gamma distribution. As for the Weibull distribution, different parametrizations for the generalized gamma distribution are present in the literature.
Due to the complexity of the hazard function, which can be obtained by diving the density function $f(t)$ through the survivor function $S(t)$, it will not be displayed here directly, but some examples for the hazard rate $\lambda(t)$ considering different parameter values are shown in Figure 2.1.
The generalized gamma distribution allows various shapes of hazard functions including decreasing and increasing patterns and covers most of the common event time distributions for certain parameter settings or as limiting distributions. The two parameter gamma distribution is obtained for ν=$\tilde{\alpha}$. With ν=1 the generalized gamma distribution translates

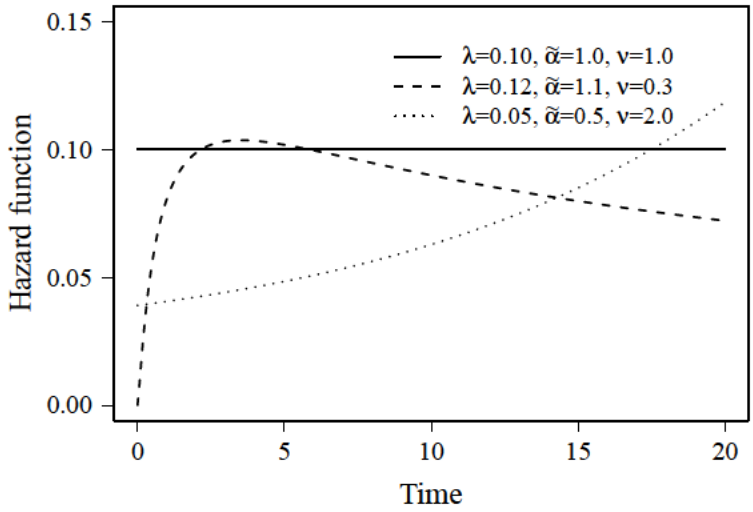

Figure 2.1: Hazard rates of the generalized gamma distribution for different choices of the parameters.

to the Weibull distribution described in Section 2.4.2 with $\alpha=1/\tilde{\alpha}$. The exponential distribution presented in Section 2.4.1 is a special case of the generalized gamma distribution with $\nu=\tilde{\alpha}=1$. The log normal distribution is a limiting case of the generalized gamma distribution for ν going to zero. A more detailed description of the generalized gamma distribution and its characteristics can be found in Cox et al (2007).
For regression purposes the influence of covariates on the parameter λ was assessed by Cox et al (2007) and Lau et al (2011) via $\lambda=\exp(-\boldsymbol{\beta}_{\lambda}^{\top}\mathbf{x})$. If enough information is available, covariate effects on the other parameters can be estimated. In the analyses presented by Cox et al (2007) a model allowing all three parameters of the generalized gamma distribution to depend on covariates was presented, which was called saturated generalized gamma model by the authors. Identity link functions were used for ν and $\tilde{\alpha}$ ($\nu=\boldsymbol{\beta}_{\nu}^{\top}\mathbf{x}$, $\tilde{\alpha}=\boldsymbol{\beta}_{\tilde{\alpha}}^{\top}\mathbf{x}$). Different authors reported on numerical problems present for parameter estimation in a generalized gamma model (Gomes et al, 2008; Noufaily and Jones, 2013).

Chapter 3

Competing risks framework

3.1 The competing risks problem

In classical time-to-event or survival analysis subjects are under risk for one terminal event. Examples are time from diagnosis of a certain disease or time from treatment initiation to patient's death in a clinical study or lifetime of a machine in quality management of a company. In some applications subjects cannot fail from just one certain type of event, but are under risk of failing from two or more mutually exclusive types of event. In a clinical study the primary endpoint may be time to death from a cardiac reason, which can be obscured by death from another reason. In a technical application the lifetime of a special component maybe of interest, which cannot be observed when the machine breaks down due to the failure of another component. When an individual is under risk of failing from K different types of event, these different event types are called competing risks (Figure 3.1). In the presence of multiple types of event standard quantities and methods used for time-to-event analysis as presented in Section 2 have to be adapted for the competing risks setting.

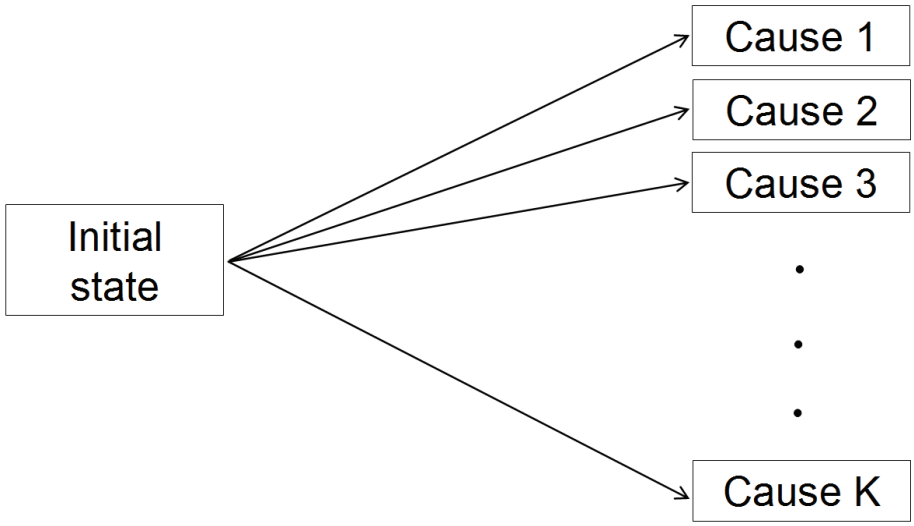

Figure 3.1: Illustration of the competing risks framework: Subjects are in an initial state at the beginning and can fail from one out of K mutually exclusive types of event.

While the competing risks problem is broadly covered in the statistical literature, which can be seen by the numbers of available textbooks (Crowder, 2001; Pintilie, 2006; Beyersmann et al, 2012), descriptions of the competing risks problem in books on survival analysis (see e.g. Marubini and Valsecchi, 1995; Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2003), and overview articles (Putter et al, 2007; Klein, 2010; Bakoyannis and Touloumi, 2011; Allignol et al, 2011), the problem is widely ignored in the medical literature, which is discussed in an article by Koller et al (2012). In that article a review of 50 clinical studies with possible competing risks endpoints, which were published in highly ranked medical journals as the *New England Journal of Medicine* or *The Lancet* from October 2007 to October 2010, was performed regarding presence of competing risks and methods used to account for competing risks. In 37 of the 50 evaluated articles (74%) at least one of the assessed endpoints implied the presence of competing risks, while the other articles only presented results on all-cause mortality or on composite endpoints not relevant for application of competing risks analysis methods. In 35 of the investigated articles inadequate analysis of competing risks data was observed, as completely ignoring the competing risks problem or performing "naïve Kaplan-Meier estimation" in the presence of competing risks, leading to biased estimates for the event probabilities (see Section 3.4). Competing risks methods were applied in only 10 of the 50 studies and only in two articles the correct estimates for the cumulative incidence function (see Section 3.3.1) were explicitly calculated and described. In summary, it can be seen that although the competing risks problem is well known and described in the statistical literature, the application of competing risks methods for analysis of clinical data is not established. In recent years, the competing risks problem appears to have become more widely recognized in the medical community, which is indicated by the publication of competing risks articles in medical journals, explaining and discussing adequate analysis methods for event time data in the presence of competing risks (Dignam and Kocherginsky, 2008; Berry et al, 2010; Roobol and Heinsdijk, 2011; Chappell, 2012; Dignam et al, 2012).
In this section different ways for presenting the competing risks framework are described (Section 3.2) and quantities used for description and analysis of competing risks data are introduced (Section 3.3). In Section 3.4 application of the standard Kaplan-Meier method in order to estimate event probabilities in the presence of competing risks, treating individuals that failed from a competing event as censored observations (a procedure that is referred to as "naïve Kaplan-Meier estimator" in the literature), is discussed.

3.2 Competing risks presentation

Two different ways of competing risks presentation can be found in the literature. Competing risks data can either be considered via a latent failure time approach, implying a joint distribution for the times to the K types of event, which is discussed in Section 3.2.1, or a competing risks process can be represented by two random variables, one random variable for the event time and one for the type of event (Section 3.2.2). Modern competing risks analyses are based on the latter approach, due to the presence of identifiability problems in the latent failure times formulation. Discussions on different approaches towards competing risks data can also be found in Pintilie (2006) and Beyersmann et al (2012).

3.2.1 Competing risks as latent failure times

One possible access to a competing risks problem is to assume, that there are random variables $T_1, \ldots, T_K$ for the time to each of the K possible event types. In a real data situation, only time to the first event, denoted as $\overline{T}$, with

$$\overline{T} = \min\{T_1, T_2, ..., T_K\}, \tag{3.1}$$

can be observed. Additionally, an indicator variable D, denoting the type of the observed event, is needed.
In order to estimate the survivor function of the joint distribution $S(t_1, \ldots, t_K)$, the correlation between times to different event types has to be assessed. In a classical competing risks setting, only one event type can be observed for each individual, so the correlation structure cannot be estimated from observable data and a correlation structure has to be assumed that cannot be verified. Tsiatis (1975) demonstrated, that for each joint distribution with assumed independence between the latent failure times, a dependence structure can be found, that provides the same likelihood. As the correlation structure cannot be estimated from observable data, Prentice et al (1978) questioned the plausibility of the latent failure times approach and therefore discouraged its use, while Beyersmann et al (2012) discussed and illustrated, that there is no gain, but additional problems, when the latent failure time approach is considered, instead of the approach presented in the next section (see Chapter 4.3.1 of Beyersmann et al, 2012).
Various analysis methods following the latent failure time approach were introduced in the literature. Cox (1959) discussed different models for analysis of exponentially distributed event types with two possible types of failure. In order to estimate plausible ranges for the joint survivor function and the marginal survivor functions for different types of event in the lack of an estimable dependence structure, several authors (Peterson, 1976; Slud and Rubinstein, 1983; Klein and Moeschberger, 1988) developed methods, which allow to derive bounds for these quantities considering possible dependence structures. In recent years approaches using copula functions to estimate the joint event time distribution from marginal distributions assuming a correlation structure between times to different types of event were presented (Kaishev et al, 2007; Lo and Wilke, 2010; Chen, 2010), which are mainly popular in the field of finance and insurance mathematics. These approaches will not be discussed in this work, due to the non-identifiability problem mentioned above and consequently the reliance on a correlation assumption, that cannot be verified using observable data.

3.2.2 Competing risks as bivariate variables

An alternative approach to competing risks is consideration of a bivariate random variable (T, D) with T being a random variable for the event time and D a random variable for the event type. The competing risks process can then be interpreted as a special case of a multistate model (see e.g. Andersen and Keiding, 2012), leading to the intuitive definitions of cumulative incidence functions and cause-specific hazard rates as presented in Sections 3.3.1 and 3.3.2. The analysis can be performed hazard-based without identifiability problems and all measures can be estimated from observable data (Prentice et al, 1978). For each individual $i = \{1, \ldots, n\}$ the couple of event time or last time known to be free of any event

t_i and a status variable indicating the type of event $d_i \in \{1, \ldots, K\}$ or a censored event time (d_i=0) is observed.
As the competing risks data are not represented by different random variables for the times to possible event types, but by one variable providing the event time and one variable indicating the type of event, the concept of statistical dependence between times to different types of event does not apply for this approach (see discussion in Chapter 7.2 of Beyersmann et al, 2012).

3.3 Important measures in the competing risks setting

As in the competing risks setting individuals can fail from different event types, measures used for standard survival analysis with only one certain type of event have to be adapted. In this section the most important and commonly used concepts and quantities are described.

3.3.1 The cumulative incidence function

In the presence of competing risks the probability for occurrence of each event type k out of the possible event types $1, ..., K$ up to a given time t can be described. That probability is mostly called "cumulative incidence function" for event type k in the literature, and is denoted as $\underline{F}_k(t)$ in this work. So

$$\underline{F}_k(t) = P(T \leq t, D = k), \tag{3.2}$$

where T is a strictly positive random variable for the event time and D is a random variable for the type of event. Some other names are existent in the statistical literature as "crude event probability" (see e.g. Tsiatis, 2005; Lambert et al, 2010) or "subdistribution function" (Resche-Rigon and Chevret, 2006; Pintilie, 2007). The name "subdistribution function" is motivated by the fact, that $\underline{F}_k(t)$ is not a real distribution function, as it does not converge to one for t going to infinity, but to the overall probability for an event of type k

$$\lim_{t \to \infty} \underline{F}_k(t) = P(D{=}k). \tag{3.3}$$

For a given time t the cumulative incidence functions of all K event types sum up to one minus the probability of being event-free up to time t, which is often called the overall survivor function and is denoted as $S_{ov.}(t)$ here

$$S_{ov.}(t) = 1 - \sum_{k=1}^{K} \underline{F}_k(t). \tag{3.4}$$

So for t going to infinity the cumulative incidence functions for all K types of event sum up to one

$$\lim_{t \to \infty} \sum_{k=1}^{K} \underline{F}_k(t) = 1. \tag{3.5}$$

In order to estimate the cumulative incidence function for event type k, the so called cause-specific hazard function, which is introduced in the next section, has to be considered. Non-parametric estimation of the cumulative incidence function is presented in Section 3.3.5. For convenience and in accordance with the majority of the statistical literature, the cumulative event probability for an event of type k up to time t, denoted as $\underline{F}_k(t)$, will be called "cumulative incidence function" throughout this work.

3.3.2 The cause-specific hazard rate

As in standard survival analysis, hazard rates play an important role for the analysis of competing risks data, as these can be estimated in the presence of censored observations. The cause-specific hazard rate for event type k is the natural adaptation of the common hazard rate shown in Equation 2.4, providing an individual's probability for failing from an event of type k in an infinitesimal small time interval t to $t+\Delta t$ given he did not fail from any event up to time t

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t, D = k \mid T \geq t)}{\Delta t}. \tag{3.6}$$

Considering mutually exclusive terminal events, the cause-specific hazards for all K event types at time t sum up to the overall hazard rate for failing from any event at t

$$\lambda_{ov.}(t) = \sum_{k=1}^{K} \lambda_k(t). \tag{3.7}$$

In analogy to standard survival analysis the cumulative cause-specific hazard rate for event type k at time t is the integral over the cause-specific hazard function from time zero to t

$$\Lambda_k(t) = \int_0^t \lambda_k(s)\mathrm{d}s. \tag{3.8}$$

The overall survivor function $S_{ov.}(t)$, denoting the probability of being free from any event up to time t, depends on the (cumulative) cause-specific hazard functions for all K types of event, which sum up to the overall (cumulative) hazard rate

$$S_{ov.}(t) = \exp\left(-\sum_{k=1}^{K} \Lambda_k(t)\right) = \exp\big(-\Lambda_{ov.}(t)\big). \tag{3.9}$$

The relationship between the cumulative incidence function for event type k and the cause-specific hazard functions can be expressed as

$$\underline{F}_k(t) = \int_0^t \lambda_k(s)\, S_{ov.}(s)\mathrm{d}s = \int_0^t \lambda_k(s)\, \exp\left(-\sum_{l=1}^{K} \Lambda_l(s)\right) \mathrm{d}s. \tag{3.10}$$

As can be seen from Equation 3.10, the cumulative incidence function for event type k depends on the cause-specific hazard functions for all K types of event, indicating that risks

for all event types have an effect on the probability for an event of type k. A consequence of that fact is, that in group comparisons a higher cause-specific hazard for an event of type k for one group does not necessarily translate to a higher cumulative incidence of that event. This issue is described in different articles (see e.g. Putter et al, 2007; Dignam and Kocherginsky, 2008; Allignol et al, 2011) and is also presented in Section 3.3.4 of this work in a discussion on differences between the cause-specific and the subdistribution hazard rate, which is introduced in the following section, and in simulated examples for hazard-based regression models in Section 4.3.

3.3.3 The subdistribution hazard rate

In order to define a "hazard-type" quantity that is directly linked to the cumulative incidence function in the presence of competing risks, Gray (1988) introduced the so called subdistribution hazard rate. The subdistribution hazard rate for event type k, denoted in this work as $\gamma_k(t)$, differs from the cause-specific hazard rate shown in Equation 3.6 by the definition of its risk set. For the subdistribution hazard rate for event type k at time t individuals that failed from an event other than k prior to t remain in the risk set

$$\gamma_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k \mid T \geq t \cup \{T < t, D \neq k\})}{\Delta t}. \tag{3.11}$$

The link between the cumulative incidence function and the subdistribution hazard is as known from standard survival analysis (see Equation 2.9)

$$\underline{F}_k(t) = 1 - \exp\big(-\Gamma_k(t)\big), \tag{3.12}$$

with $\Gamma_k(t)$ denoting the cumulative subdistribution hazard

$$\Gamma_k(t) = \int_0^t \gamma_k(s)\mathrm{d}s. \tag{3.13}$$

Competing events do not have to be accounted for explicitly, as these are considered implicitly in the adapted risk set. As $\gamma_k(t)$ provides the properties of a hazard rate for the subdistribution function $\underline{F}_k(t)$, it is called subdistribution hazard.
Due to its direct relationship to the cumulative incidence function, the subdistribution hazard became very popular in recent years. Different methods focussing on the subdistribution hazard were proposed, as the widely used proportional subdistribution hazards regression model introduced by Fine and Gray (1999), which is described in Section 4.2.
Some authors do not use an index for the subdistribution hazard, but only use it for the event of interest, as the estimation of subdistribution hazards for different event types is under discussion due to the risk set definition (see e.g. Beyersmann et al, 2012). For convenience, the index for the type of event is denoted here for the subdistribution hazard. Andersen and Keiding (2012) questioned the usefulness and the interpretability of the subdistribution hazard due to the unintuitive procedure of keeping individuals that failed from a competing event in the risk set for later timepoints.

3.3.4 Relationship between cause-specific and subdistribution hazard rate

The relationship between the cause-specific and the subdistribution hazard rate can be derived analytically via the relationships to the cumulative incidence function shown in Equations 3.10 and 3.12 (Beyersmann and Schumacher, 2007). A detailed derivation of that relationship is presented by Beyersmann et al (2012). In the case of two possible endpoints

$$\lambda_1(t) = \gamma_1(t) \left(1 + \frac{\underline{F}_2(t)}{S_{ov.}(t)}\right), \tag{3.14}$$

with $\lambda_1(t)$ denoting the cause-specific hazard for the event of interest (k=1), $\gamma_1(t)$ the corresponding subdistribution hazard, $\underline{F}_2(t)$ the cumulative incidence function for the competing event (k=2), and $S_{ov.}(t)$ the overall survivor function, providing the probability of freedom from any event up to time t. As can be seen from Equation 3.14, the subdistribution hazard for event type k=1 is related to the cause-specific hazards of both event types, as the cumulative incidence function for event type k=2 and the overall survivor function depend on the cause-specific hazards for both types of event. Therefore, analysis of the cause-specific and the subdistribution hazards will generally lead to different results in the presence of competing risks.
In Figure 3.2 the cause-specific and the subdistribution hazard for an event of interest are shown for various values of the cause-specific hazard for the competing event. For all scenarios the cause-specific hazard for the event of interest was chosen to be λ_1=0.10. The cause-specific hazard for the competing event, which is indicated at the top of each picture, was chosen to be:

a) $\lambda_2 = 0.01$
b) $\lambda_2 = 0.05$
c) $\lambda_2 = 0.10$
d) $\lambda_2 = 0.25$

Figure 3.2 reveals, that the difference between cause-specific and subdistribution hazard depends on the risk for a competing event, which is driven by the cause-specific hazard $\lambda_2(t)$. It follows from Equation 3.14 and from definition of the risk set in Equation 3.11 that the cause-specific and the subdistribution hazard are equal in the absence of competing risks, i.e. in the standard survival setting with one possible endpoint, and that they have to approach the same value for t going to zero in the presence of competing risks. From Equation 3.12 follows, that the subdistribution hazard has to converge to zero for t going to infinity, as the cumulative incidence function approaches a value smaller than one in the presence of competing risks, and therefore the cumulative subdistribution hazard function has to converge to a finite value.

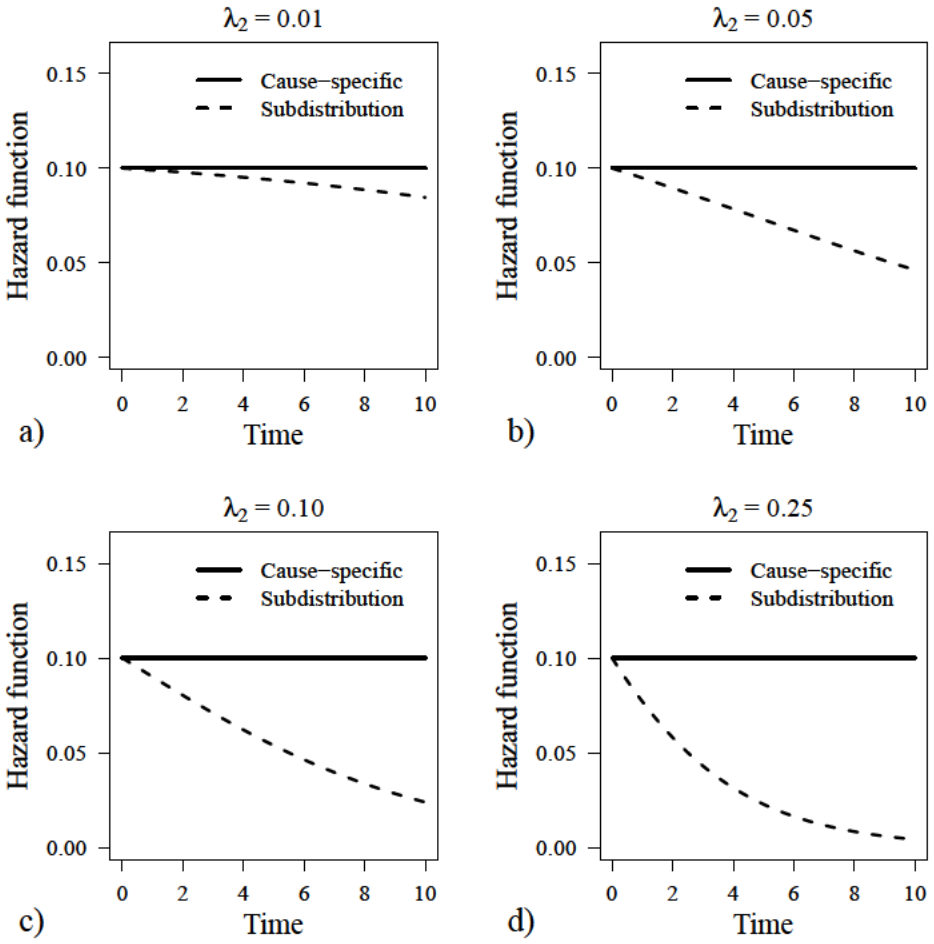

Figure 3.2: Illustration of the relationship between the cause-specific and the subdistribution hazard for the event of interest in dependence of the cause-specific hazard for the competing event. The cause-specific hazard for the event of interest was chosen to be 0.10, the cause-specific hazard for the competing event 0.01, 0.05, 0.10, and 0.25, as indicated at the top of each picture.

3.3.5 Nonparametric estimators

Nonparametric estimators for the basic quantities introduced above are described in this section. More detailed descriptions and derivations of the presented formulas can be found in Putter et al (2007) or Beyersmann et al (2012).

Estimating the cause-specific hazard rate

Nonparametric estimation of the cause-specific hazard function for event type k is analogous to estimation of the hazard function in a standard survival setting (see Equation 2.5), with the difference that only events of interest, here of type k, are considered. With $\tilde{\mathbf{t}}_{\mathbf{k}}=(\tilde{t}_{k1},\ldots,\tilde{t}_{kN_k})$ being the vector of observed event times with events of type k, $d_{k\tilde{t}_{ki}}$ denoting the number of observed events of type k at time $\tilde{t}_{ki}$, and $R_{\tilde{t}_{ki}}$ the number of individuals at risk at $\tilde{t}_{ki}$, i.e. the number of subjects that did not experience any event $1,\ldots,K$ before time $\tilde{t}_{ki}$ and that are still under observation at $\tilde{t}_{ki}$, the estimate for the

cause-specific hazard at $\tilde{t}_{ki}$ is

$$\hat{\lambda}_k(\tilde{t}_{ki}) = \frac{d_{k\tilde{t}_{ki}}}{R_{\tilde{t}_{ki}}}. \tag{3.15}$$

For timepoints without an observed event of type k, the nonparametric estimate of the cause-specific hazard for event type k is zero.
The cumulative cause-specific hazard rate $\Lambda_k(t)$ can be estimated analogously to Equation 2.7 considering only events of type k. The Nelson-Aalen estimator for the cumulative cause-specific hazard function can be derived using the R package *mvna* (Allignol et al, 2008).

Estimating the subdistribution hazard rate

The subdistribution hazard rate can only be estimated in a similar way to the cause-specific hazard as shown in Equation 3.15, when no censored observations are present. When an event was observed for each subject, the subdistribution hazard at a timepoint $\tilde{t}_{ki}$ with an observed event of type k can be estimated as

$$\hat{\gamma}_k(\tilde{t}_{ki}) = \frac{d_{k\tilde{t}_{ki}}}{R^*_{\tilde{t}_{ki}}}, \tag{3.16}$$

where $d_{k\tilde{t}_{ki}}$ again denotes the number of type k failures at $\tilde{t}_{ki}$ and $R^*_{\tilde{t}_{ki}}$ is the adapted risk set, including all subjects that did not fail from any of the K events up to time $\tilde{t}_{ki}$ and all subjects that failed from an event other than k before $\tilde{t}_{ki}$ (see definition of the subdistribution hazard in Equation 3.11). For all timepoints without an observed event of type k, the nonparametric estimator for the subdistribution hazard returns zero.
In the presence of censored observations a potential censoring time has to be derived for patients that failed from an event other than k, in order to obtain an unbiased estimate for the subdistribution hazard for event type k. Estimation of the potential censoring time is described in Gray (1988) and Fine and Gray (1999) for different scenarios of administrative or non-administrative censoring. The estimation procedure in the presence of censored observations will be discussed in more detail for the subdistribution hazards regression presented in Section 4.2.

Estimating the cumulative incidence function

In order to estimate the cumulative incidence function, the overall survivor function shown in Equation 3.9 and the cause-specific hazard rate for the event of interest k have to be estimated in a first step. The overall survivor function can be estimated using the Kaplan-Meier estimator (Kaplan and Meier, 1958) shown in Equation 2.3 treating failures from any cause as events. An estimate for the cause-specific hazard rate can be obtained as described in Equation 3.15. An estimate for the cumulative incidence function for event k can be derived from estimates of these two measures using the following equation (see e.g. Putter et al, 2007), assuming an ordered vector of event times with events of type k,

$$\underline{\hat{F}}_k(t) = \sum_{i:\, \tilde{t}_{ki} \leq t} \hat{\lambda}_k(\tilde{t}_{ki})\, \hat{S}_{ov.}(\tilde{t}_{k(i-1)}). \tag{3.17}$$

A step function is returned by the estimator for the cumulative incidence function, with jumps at timepoints with an observed event of type k, and constant values for timepoints without an observed event or with an observed competing event.
That estimator for the cumulative incidence function in a competing risks setting is a special case of the Aalen-Johansen estimator for transition probabilities in multi-state models (Aalen, 1978b). Using that equation, cumulative incidence functions for all K types of event can be estimated separately. Different proposals for variance estimators for the cumulative incidence function, either based on martingale theory (Pepe, 1991; Korn and Dorey, 1992; Lin, 1997) or the multinomial normal distribution (Gaynor et al, 1993; Betensky and Schoenfeld, 2001), were presented in the literature. A comparison and discussion of different variance estimators can be found in Braun and Yuan (2007) with a special focus on small sample performances.

3.3.6 Illustrative example: Comparison of cause-specific and subdistribution hazard estimates

In Figure 3.3 nonparametric estimates of the cause-specific and the subdistribution hazard rates (a) and the cumulative cause-specific and subdistribution hazard rates (b) are compared for a simulated data example without censored observations, in order to illustrate differences between these quantities. The observed event times and the the risk sets, needed for estimation of the cause-specific and the subdistribution hazards, as well as the estimated hazard rates are shown in Table 3.1.

i	t_i	d_i	$\tilde{t}_{1i}$	$d_{1\tilde{t}_{1i}}$	$R_{\tilde{t}_{1i}}$	$R^*_{\tilde{t}_{1i}}$	$\hat{\lambda}_1(t_i)$	$\hat{\gamma}_1(t_i)$	$\hat{\Lambda}_1(t_i)$	$\hat{\Gamma}_1(t_i)$
1	2	1	2	1	7	7	$\frac{1}{7} = 0.14$	$\frac{1}{7} = 0.14$	0.14	0.14
2	7	1	7	1	6	6	$\frac{1}{6} = 0.17$	$\frac{1}{6} = 0.17$	0.31	0.31
3	12	2	—	—	—	—	—	—	0.31	0.31
4	14	1	14	1	4	5	$\frac{1}{4} = 0.25$	$\frac{1}{5} = 0.20$	0.56	0.51
5	18	2	—	—	—	—	—	—	0.56	0.51
6	23	1	23	1	2	4	$\frac{1}{2} = 0.50$	$\frac{1}{4} = 0.25$	1.06	0.76
7	30	2			—	—	—	—	1.06	0.76

Table 3.1: Fictive example data to illustrate differences in the estimation of cause-specific and subdistribution hazards. Estimated hazard rates and cumulative hazard rates are also shown in Figure 3.3.

The table shows:

- i: Index for the ordered event times
- t_i: Observed event time
- d_i: Observed type of event
- $\tilde{t}_{1i}$: Observed event times with an event of type k=1
- $d_{1\tilde{t}_{1i}}$: Number of events of type k=1 at $\tilde{t}_{1i}$

- $R_{\tilde{t}_{1i}}$: Risk set for estimation of the cause-specific hazard for k=1 at $\tilde{t}_{1i}$
- $R^*_{\tilde{t}_{1i}}$: Risk set for estimation of the subdistribution hazard for k=1 at $\tilde{t}_{1i}$
- $\hat{\lambda}_1(t_i)$: Estimated cause-specific hazard for k=1 at t_i
- $\hat{\gamma}_1(t_i)$: Estimated subdistribution hazard for k=1 at t_i
- $\hat{\Lambda}_1(t_i)$: Estimated cumulative cause-specific hazard for k=1 at t_i
- $\hat{\Gamma}_1(t_i)$: Estimated cumulative subdistribution hazard for k=1 at t_i

As individuals, that failed from a competing event, remain in the risk set for the subdistribution hazard, the estimated subdistribution hazard for the event of interest will be smaller than the estimated cause-specific hazard after occurrence of the first competing event, with the difference between the both hazard functions depending on the amount of competing events as illustrated in Figure 3.2.

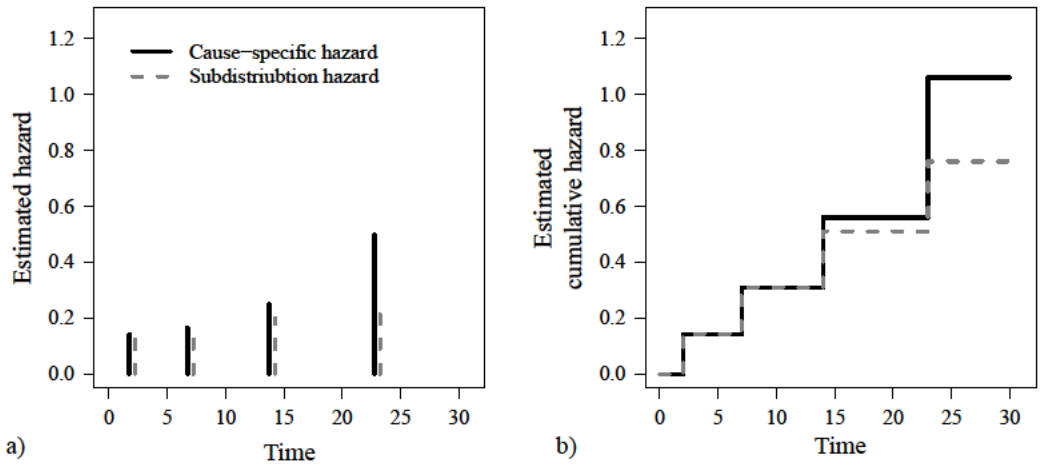

Figure 3.3: Illustration of differences in estimates of the cause-specific and subdistribution hazard rates (left picture) and consequences on estimated cumulative hazard rates (right picture) in completely observed data shown in Table 3.1.

3.4 The naïve Kaplan-Meier estimator

Often analysis of competing risks data, especially in the medical literature, is performed by application of the standard Kaplan-Meier estimator treating competing events as censored observations. This procedure, which is sometimes called the "naïve Kaplan-Meier estimator" in the literature (see e.g. Putter et al, 2007), violates one important assumption of the standard Kaplan-Meier estimator, namely the independence of event times and censoring times. Individuals that failed from a competing event cannot fail from the event of interest later, but as they are treated like individuals that are event-free and not followed any further, the same risk for a later event of interest is allocated to these individuals. Using the naïve Kaplan-Meier estimator, the probability for an event of type k up to a given time t is estimated by

$$1 - \hat{S}_k(t) = \exp\big(-\hat{\Lambda}_k(t)\big), \tag{3.18}$$

where $S_k(t)$ can be estimated from the observed data, but cannot be interpreted as a marginal survival probability in the "real" world (see discussion in Putter et al, 2007). The naïve Kaplan-Meier estimator overestimates the true event probability in the presence of competing risks, which can be seen in a comparison to the correct formula for estimation of the cumulative incidence function,

$$\begin{aligned} 1 - S_k(t) &= \int_0^t \lambda_k(s) \exp\big(-\Lambda_k(s)\big) \mathrm{d}s \\ &\geq \int_0^t \lambda_k(s) \exp\left(- \sum_{l=1}^{K} \Lambda_l(s) \right) \mathrm{d}s = \underline{F}_k(t) \end{aligned} \tag{3.19}$$

with equality only valid, if $\lambda_l(t)=0$ for all $l \neq k$ and all t, i.e. in the absence of competing risks. In an example presented by Putter et al (2007) application of the naïve Kaplan-Meier estimator lead to cumulative incidence functions for the two possible event types that summed up to a value larger than one. This is shown here using data from a study published by Essler et al (2013), investigating the time from the beginning of a hypofractionated stereotactic body radiation therapy (SBRT) to either tumour-related death or death from another cause in a population of 29 patients suffering from non-small cell lung cancer (NSCLC) of stage I, that were not suitable for surgery. In the left column of Figure 3.4 results of the naïve Kaplan-Meier estimator are presented for probabilities of tumour-related death (first row) and death from other causes (second row). These lead to an estimate for the probability of death from any cause, which was calculated as the sum of the two former event probabilities, that was greater than one after 40 months of follow-up, as competing risks were not considered adequately for estimation of the cumulative event probabilities. In the right column of Figure 3.4 the according cumulative incidence functions were estimated as shown in Equation 3.17, so for each timepoint t one minus the sum of the cumulative incidence functions provides the value of the Kaplan-Meier estimator for overall survival, which is shown in the picture in the bottom row of the right column.
In the presence of independence between the event times to different types of event and independence between the event time distribution and the censoring distribution, assumptions that cannot be checked from observable data (see discussion on the latent failure time approach in Section 3.2.1), $S_k(t)$ can be interpreted as survivor function in a hypothetical world, where the competing events were eliminated. The applicability and usefulness of the naïve Kaplan-Meier estimator and its interpretation as marginal event time distribution in a hypothetical world without the competing events is widely discussed in statistical and medical literature. This can be seen in a discussion on an Editorial Letter published by Bodnar and Blackstone (2005) in the *Journal of Heart Valve Disease* encouraging application of the Kaplan-Meier estimator for evaluation of the usefulness of tissue valves, assessed by time to valve failure with death before valve failure as competing event. A Letter to the Editor (Grunkemeier et al, 2006) questioning that approach was published later in the same journal as well as an answer of the editors that still argued for the use of the Kaplan-Meier estimator in that situation (Bodnar and Blackstone, 2006). A full article on the topic was published one year later by Grunkemeier et al (2007) in another journal intended for the same audience.

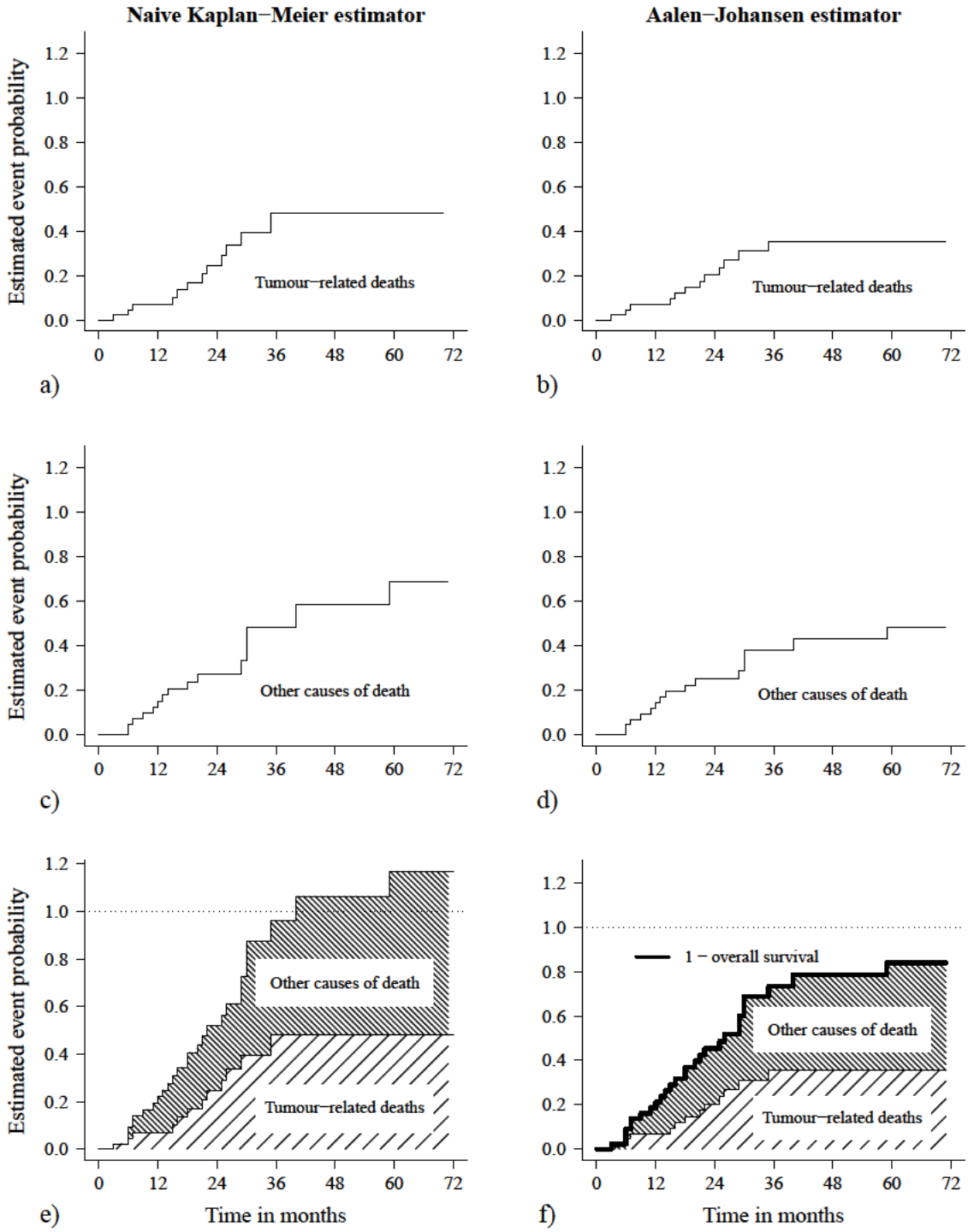

Figure 3.4: Left column: Estimates of probabilities for tumour-related death (top) and deaths from other causes (middle) using the naïve Kaplan-Meier estimator and the sum of both estimates (bottom). Right column: Estimates of cumulative incidence functions for tumour-related death (top) and deaths from other causes (middle) and the sum of both estimates using the Aalen-Johansen estimator, providing an adequate estimate for one minus overall survival (bottom).

From a biological standpoint, some authors question the independence assumption in clinical settings (Moeschberger and Klein, 1995; Crowder, 2001) and therefore discourage the use of the Kaplan-Meier estimator in the presence of competing risks with an unknown dependence structure. Andersen and Keiding (2012) recommend to "stick to this world" for analysis and interpretation of competing risks data, and therefore argue against an interpretation of the Kaplan-Meier estimate for a hypothetical world, but recommend presentation of cause-specific hazard rates and cumulative incidence functions.
Regardless of the interpretability of the Kaplan-Meier estimator as an estimator for the marginal event time distribution in a hypothetical world, in most publications results from the naïve Kaplan-Meier estimator are presented as an estimate for one minus the event probability without discussion of hypothetically extinguishing competing events and without questioning the independence assumption. This procedure returns biased estimates for one minus the cumulative incidence function, since competing events are not considered adequately, as discussed above.

Chapter 4

Regression approaches for the competing risks setting

Since the end of the 1970s several regression approaches for the competing risks setting were introduced. The most commonly used approaches are the cause-specific hazards regression proposed by Prentice et al (1978) and the subdistribution hazards regression introduced by Fine and Gray (1999). Two other approaches are based on the factorisation of the joint distribution of event times and event types into one marginal and one conditional distribution. Larson and Dinse (1985) considered the product of the marginal event type distribution and the conditional distribution of event times given the type of event. The approach was later extended amongst others by Ng and McLachlan (2003), Escarela and Bowater (2008), and Lau et al (2008). In the so called vertical modelling approach Nicolaie et al (2010) proposed to use the product of the marginal event time distribution and the conditional distribution of event types given the time of event, in order to obtain estimates for relative hazard rates over the course of time.
In this section these regression approaches are described. Additionally, a computation technique using pseudo-observations, which was introduced by Andersen et al (2003) and Klein and Andersen (2005) for regression purposes in a competing risks setting, is sketched. A summary of the different approaches with an application on a clinical data set was published (Haller et al, 2013). Results of the data application are also presented in Section 8 of this work.

4.1 Cause-specific hazards regression

In the competing risks setting, as in common survival analysis, a measure of interest that can be used in the presence of censored observation has to be considered. Prentice et al (1978) proposed to estimate the effect of covariates on the cause-specific hazard rates. The cause-specific hazards approach is appealing as these "completely determine the competing risks process" (Beyersmann et al, 2009) and parameters can be estimated from observable data using standard software.

4.1.1 Estimation of regression coefficients

Following the notation by Prentice et al (1978), for each individual i the data $(t_i, j_i, \delta_i, \mathbf{x_i})$ are observed, where t_i is the observed time, j_i is the observed cause of failure, δ_i is a censoring indicator returning the value of zero for a censored observation and a value of one if any event was observed, and $\mathbf{x_i}$ is the vector of covariates, which is assumed to be constant over time. For a censored observation an arbitrary value can be set for j_i. The likelihood function under independent censoring can be written as

$$\begin{aligned} L &= \prod_{i=1}^{n} \Big(\lambda_{j_i}(t_i|\mathbf{x_i})^{\delta_i} \, S(t_i|\mathbf{x_i}) \Big) = \\ &= \prod_{i=1}^{n} \left(\lambda_{j_i}(t_i|\mathbf{x_i})^{\delta_i} \prod_{l=1}^{K} \exp\Big(- \int_0^{t_i} \lambda_l(s|\mathbf{x_i}) \mathrm{d}s \Big) \right), \end{aligned} \tag{4.1}$$

which is an adaptation of the likelihood function used in standard survival analysis (see Equation 2.16) considering the relationship between the overall survivor function and the cause specific hazards shown in Equation 3.9.
Using the representation of competing risks data and covariates as a triple $(t_i, d_i, \mathbf{x_i})$ with d_i indicating the type of event ($d_i \in \{1, \ldots, K\}$) or a censored observation (d_i=0), the likelihood function can be written equivalently as

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\Big(\prod_{l=1}^{K} \lambda_l(t_i|\mathbf{x_i})^{I(d_i=l)} \Big) \, S(t_i|\mathbf{x_i}) \right) = \\ &= \prod_{i=1}^{n} \left(\Big(\prod_{l=1}^{K} \lambda_l(t_i|\mathbf{x_i})^{I(d_i=l)} \Big) \prod_{l=1}^{K} \exp\Big(- \int_0^{t_i} \lambda_l(s|\mathbf{x_i}) \mathrm{d}s \Big) \right). \end{aligned} \tag{4.2}$$

The form of the likelihood function presented in Equations 4.1 or 4.2 leads to some important implications, which are further discussed by Prentice et al (1978):

- The hazard functions and the regression coefficients are identifiable and can be estimated from the observed data.
- The score function for estimation of regression coefficients for the event of interest does not change, when all observed competing events are treated like censored observations. Therefore, standard methods for estimation of hazard rates or hazard ratios can be applied treating competing events as censored observations.
- Covariate effects on cause-specific hazards for different event types can be estimated in separate regression models.

When regression models for all event types are fit to the data in order to model the complete competing risks process, different sets of covariates might be considered for different types of event, denoted by an according index. Throughout this work it is assumed, that the set of covariates is the same for all K types of event.
Using this regression model allows estimation of covariate effects on the cause-specific

hazards. Due to dependence of the cumulative incidence function on the cause-specific hazards for all possible types of event (see Equation 3.10), an effect on the cause-specific hazard of a certain event type does not necessarily translate into an effect on the event probability, represented by the cumulative incidence function. This fact is further discussed and illustrated in Section 4.1.2, describing how the cumulative incidence function can be estimated from proportional cause-specific hazards regression models for a given vector of covariates, and in Section 4.3 discussing differences between the cause-specific and the subdistribution hazards regression model, which is presented in Section 4.2.
Prentice et al (1978) proposed to consider a Cox-type regression model (Cox, 1972) to estimate covariate effects on the cause-specific hazard rates, assuming proportional cause-specific hazards

$$\lambda_k(t|\mathbf{x}) = \lambda_{k;0}(t)\, \exp(\boldsymbol{\beta_k}^\top \mathbf{x}). \tag{4.3}$$

Here $\lambda_{k,0}(t)$ describes the cause-specific baseline hazard for event type k, which is considered as high-dimensional nuisance parameter, when covariate effects are estimated, $\mathbf{x}$ is the P-dimensional vector of covariates and $\boldsymbol{\beta_k}$ is the vector of regression coefficients of length P for the k^{th} type of event.
The vector $\tilde{\boldsymbol{t}}_{\boldsymbol{k}} = (\tilde{t}_{k1}, \ldots, \tilde{t}_{kN_k})$ represents the N_k observed failure times with an event of type k. It is assumed, that at each of these timepoints one individual fails from an event of type k. Methods for handling tied event times are e.g. discussed in Therneau and Grambsch (2000).
When a Cox proportional hazards model is used to assess the covariate effects on the cause-specific hazards, the partial likelihood can be denoted as

$$PL = \prod_{k=1}^{K} \prod_{i=1}^{N_k} \frac{\exp(\boldsymbol{\beta_k}^\top \mathbf{x_{(i)}})}{\sum\limits_{j \in R_{\tilde{t}_{ki}}} \exp(\boldsymbol{\beta_k}^\top \mathbf{x_j})}, \tag{4.4}$$

with $\mathbf{x_{(i)}}$ being the vector of regression coefficients of the individual failing from event k at time $\tilde{t}_{ki}$ and $R_{\tilde{t}_{ki}}$ denoting the risk set at $\tilde{t}_{ki}$. Due to the factorization of the partial likelihood, the regression coefficients for the different types of event can be estimated from separate models, if no common effects or baseline hazards are assumed. Estimation of regression coefficients can be conducted numerically using a Newton-Raphson algorithm. The regression coefficients $\beta_{k,1}, \ldots, \beta_{k,P}$, can be interpreted as cause-specific log-hazard ratios for event type k.
In order to estimate the cumulative incidence function from a Cox model for the cause-specific hazards, the cause-specific baseline hazard functions $\lambda_{k;0}(t)$ have to be estimated for all event types $k = \{1, \ldots K\}$. Marubini and Valsecchi (1995) present a generalized version of the Breslow estimator (Breslow, 1972) for the baseline hazard in a proportional hazards model

$$\hat{\lambda}_{k;0}(\tilde{t}_{ki}) = \frac{d_{k\tilde{t}_{ki}}}{\sum\limits_{j \in R_{\tilde{t}_{ki}}} \exp(\hat{\boldsymbol{\beta}}_k^\top \mathbf{x_j})}, \tag{4.5}$$

where $R_{\tilde{t}_{ki}}$ describes the risk set at time $\tilde{t}_{ki}$, $d_{k\tilde{t}_{ki}}$ is the number of events of type k at time $\tilde{t}_{ki}$ and $\hat{\boldsymbol{\beta}}_k$ is the estimate for the vector of regression coefficients for event type k from a proportional cause-specific hazards regression model. For each timepoint t without

an observed event of type k, the estimate for the cause-specific baseline hazard is zero. Consequently, the cumulative cause-specific baseline hazard rate for event type k can be estimated as

$$\hat{\Lambda}_{k;0}(t) = \sum_{i:\, \tilde{t}_{ki} \leq t} \frac{d_{k\tilde{t}_{ki}}}{\sum\limits_{j \in R_{\tilde{t}_{ki}}} \exp(\hat{\boldsymbol{\beta}}_k^\top \mathbf{x_j})}. \tag{4.6}$$

4.1.2 Predicting the cumulative incidence function

The cumulative incidence function for a certain type of event can be estimated according to Equation 3.17, using the Aalen-Johansen estimator under consideration of the covariate information. Assuming the vector of event times with an observed event of type k, denoted as $\tilde{\mathbf{t}}_\mathbf{k} = (\tilde{t}_{k1}, \ldots, \tilde{t}_{kN_k})$, to be ordered, the estimator for the cumulative incidence function of event type k can be written as

$$\begin{aligned}
\underline{\hat{F}}_k(t|\mathbf{x}) &= \sum_{i:\, \tilde{t}_{ki} \leq t} \hat{\lambda}_k(\tilde{t}_{ki}|\mathbf{x})\, \hat{S}(\tilde{t}_{k(i-1)}|\mathbf{x}) \\
&= \sum_{i:\, \tilde{t}_{ki} \leq t} \hat{\lambda}_{k;0}(\tilde{t}_{ki}) \exp(\hat{\boldsymbol{\beta}}_k^\top \mathbf{x})\, \exp\left(- \sum_{l=1}^{K} \hat{\Lambda}_l(\tilde{t}_{k(i-1)}|\mathbf{x}) \right) \\
&= \sum_{i:\, \tilde{t}_{ki} \leq t} \hat{\lambda}_{k;0}(\tilde{t}_{ki}) \exp(\hat{\boldsymbol{\beta}}_k^\top \mathbf{x})\, \exp\left(- \sum_{l=1}^{K} \hat{\Lambda}_{l;0}(\tilde{t}_{k(i-1)})\, \exp(\hat{\boldsymbol{\beta}}_l^\top \mathbf{x}) \right).
\end{aligned} \tag{4.7}$$

While competing events can be treated like censored observations for the estimation of cause-specific hazard rates, competing events have to be considered adequately for the estimation of cumulative incidence functions. As can be seen in Equation 4.7, the cumulative incidence function for event type k depends on the cause-specific hazards of all event types, as previously discussed in Section 3.3.2. Therefore, an observed effect on the cause-specific hazard does not necessarily translate into an effect on the cumulative incidence function. This is further discussed in Section 4.3.

4.1.3 Extensions - further reading

A proportional cause-specific hazards model was presented here, as the model originally introduced by Cox (1972) for common event time analysis is well known and the most frequently used regression model for standard survival analysis as well as for competing risks analysis. Nevertheless, a variety of other approaches or extensions were proposed in the literature.

Lunn and McNeil (1995) presented data duplication methods allowing joint estimation of regression coefficients in Cox-type cause-specific hazards regression models for different types of event and to test for differences in regression coefficients for different event types using standard survival software. An extension of the Cox regression model considering and testing time-dependence of covariate effects was proposed by Sun et al (2008). Belot et al (2010) proposed to use smooth cubic regression splines in order to obtain flexible models allowing different shapes of cause-specific baseline hazards for different event types and time-dependent covariate effects.

4.1.4 Available software

The approach presented above, estimating regression coefficients for different event types from separate models, is easily applicable using standard survival software, treating competing events as censored observations. This can be obtained using the *coxph* function of the *survival* package in R (Therneau, 2011) or *PROC PHREG* in SAS. Standard methods to check model assumptions, as plotting Schoenfeld residuals (Schoenfeld, 1982) or Martingale residuals (see e.g Therneau and Grambsch, 2000), can be conducted. The analysis of competing risks data focussing on the cause-specific hazards can also be performed using the *timereg* package in R. The use of the package in a competing risks setting is described in Scheike and Zhang (2011).
For comparison of cause-specific hazard rates between a discrete number of groups without further covariate adjustment, the standard logrank test, treating competing events as censored observations, can be applied.

4.2 Subdistribution hazards regression

In 1999 Fine and Gray developed a regression model for time-to-event data in the presence of competing risks, that focusses on the subdistribution hazard rate. It is known under the name Fine and Gray regression model or also – which might be misleading – under the name competing risks regression.
In their original article published in 1999 Fine and Gray proposed to use a Cox-type regression model for the subdistribution hazard for an event of interest, here k=1, assuming proportional subdistribution hazard rates

$$\gamma_1(t|\mathbf{x}) = \gamma_{1;0}(t)\ \exp(\boldsymbol{\eta_1}^\top \mathbf{x}). \tag{4.8}$$

$\gamma_1(t|\mathbf{x})$ denotes the subdistribution hazard for the event of interest depending on the vector of covariates $\mathbf{x}$, $\gamma_{1;0}(t)$ is the baseline subdistribution hazard for a (possibly fictitious) individual with all covariates equalling zero, and $\boldsymbol{\eta_1}$ is the vector of regression coefficients. As the competing events are incorporated implicitly in the adapted risk set (see Section 3.3.3) only a model for the event of interest k=1 is presented. In general, the proportionality assumption cannot hold true for separate subdistribution hazards regression models for different types of event (see e.g. the discussion in Chapter 5.3.4 of Beyersmann et al, 2012). Grambauer et al (2010) investigated the impact of model misspecification. They demonstrated that a subdistribuiton hazards regression model has a proper interpretation, even when the subdistribution hazards were falsely assumed to be proportional. The estimated regression coefficients can be interpreted as average subdistribution log-hazard ratios. It must be considered, that in this case the average subdistribution hazard ratio will depend on the length of follow-up (see e.g. Schemper et al, 2009).

4.2.1 Estimation of regression coefficients

For estimation of the regression coefficients in a subdistribution hazards regression model a different risk set is needed than for to the cause-specific hazards regression model described

in Section 4.1. While estimation of the regression coefficients is straightforward when complete data are observed for all individuals and under administrative censoring, the estimating procedure becomes more complicated for incomplete data with non-administrative censoring, in order to obtain unbiased estimates. As in the original article by Fine and Gray (1999), the different scenarios are described separately.

Completely observed data

When complete data are available, i.e. event time and type of event were observed for each individual, the likelihood for the proportional subdistribution hazards regression model can be written as shown in Equation 4.4, but an adapted risk set $R^*_{\tilde{t}_{ki}}$ is considered

$$R^*_{\tilde{t}_{ki}} = \{j\colon (t_j \geq \tilde{t}_{ki}) \cup (t_j \leq \tilde{t}_{ki} \cap d_j \neq k)\}, \tag{4.9}$$

including all individuals that are still under observation at $\tilde{t}_{ki}$, that means all individuals, who have not failed from any cause before $\tilde{t}_{ki}$, and all individuals, that have failed from an event other than k before time $\tilde{t}_{ki}$. Estimation of regression coefficients can be conducted as described for the cause-specific hazards regression, but using the adapted risk set. The baseline subdistribution hazard can be estimated as described in Equation 4.5 also using the adapted risk set, additionally including individuals that have failed from a competing event before the timepoint under investigation.

Administrative censoring

Administrative censoring, which is called "censoring complete data" in the article by Fine and Gray, means that individuals are only censored when they are still alive at the end of the study, but no drop-outs or losses to follow-up are present. As the maximum follow-up time is known for each individual at trial allocation, the potential censoring time c is also known for individuals that experienced any event during the trial. When only administrative censoring is present, the risk set used for parameter estimation is defined as

$$R^*_{\tilde{t}_{ki}} = \{j\colon (\min(c_j, t_j) \geq \tilde{t}_{ki}) \cup (t_j \leq \tilde{t}_{ki} \cap d_j \neq k \cap c_j \geq \tilde{t}_{ki})\}. \tag{4.10}$$

An individual j that either died from an event of interest (k=1) before the timepoint under investigation $\tilde{t}_{ki}$, was censored event-free before $\tilde{t}_{ki}$, or failed from a competing event before $\tilde{t}_{ki}$, but has a potential censoring time c_j smaller than $\tilde{t}_{ki}$, is removed from the risk set. So the risk set $R_{\tilde{t}_{ki}}$ consists of all individuals that are still under observation at $\tilde{t}_{ki}$ or that failed from an event other than k before $\tilde{t}_{ki}$, but have a potential follow-up time larger than $\tilde{t}_{ki}$.

Incomplete data

In the presence of incomplete data, i.e. when individuals dropped out of the study or were lost to follow-up (possibly additionally to administrative censoring), Fine and Gray proposed to use a weighted score function for parameter estimation, in order to obtain unbiased estimates for the regression coefficients $\boldsymbol{\eta_1}$. For construction of the score function the inverse probability of censoring weighting (IPCW) approach introduced by Robins and

Rotnitzky (1992) is used. The score function, that is maximized in order to obtain the maximum partial likelihood estimates, is weighted using time-dependent weights based on the Kaplan-Meier estimates for the survivor function of the censoring distribution. Each individual i contributes to the score function with the weight

$$w_i(t) = r_i(t)\frac{\hat{G}(t)}{\hat{G}\big(min(t_i, t)\big)}, \tag{4.11}$$

with $r_i(t)$ indicating knowledge of the vital status of individual i at time t, i.e. $r_i(t)$ is one, if individual i is known to be alive at time t or if it is known that individual i had failed before t from any cause of failure, and $r_i(t)$ is zero, if individual i was censored before time t. $\hat{G}(t)$ is the Kaplan-Meier estimate for the survivor function of the censoring time distribution at time t, that is derived from observable data. The weighted contribution of each individual is also presented and discussed in Beyersmann et al (2012).
Fine and Gray (1999) derived that their presented procedure provides consistent estimates for the subdistribution log-hazard ratios. Formulation and derivation of the variance-covariance matrix can also be found there.

4.2.2 Predicting the cumulative incidence function

The predicted cumulative incidence function for a given vector of covariates $\mathbf{x}$ can be obtained from the estimated regression coefficients using the relationship between the subdistribution hazard and the cumulative incidence function (Equation 3.12) without further consideration of effects on the competing events

$$\begin{aligned}\hat{\underline{F}}_1(t|\mathbf{x}) &= 1 - \exp\big(-\hat{\Gamma}_1(t|\mathbf{x})\big) \\ &= 1 - \exp\left(-\int_0^t \hat{\gamma}_1(s|\mathbf{x})\mathrm{d}s\right) \\ &= 1 - \exp\left(-\int_0^t \hat{\gamma}_{1;0}(s)\, \exp(\hat{\boldsymbol{\eta}}_\mathbf{1}^\top \mathbf{x})\mathrm{d}s\right).\end{aligned} \tag{4.12}$$

Estimation of a confidence band for the cumulative incidence function derived from a proportional subdistribution hazards model is described in detail in Fine and Gray (1999).

4.2.3 Extensions - further reading

In recent years different extensions of the proportional subdistribution hazards model were introduced. Latouche et al (2005) and Beyersmann and Schumacher (2008) investigated and discussed the incorporation of time-dependent covariates into a subdistribution hazards regression model. A frailty subdistribution hazards regression model was introduced by Katsahian et al (2006) in order to deal with clinical data collected in a multicentre study assuming a random centre effect. Another subdistribution hazards model for correlated data was proposed later by Dixon et al (2011). Geskus (2011) and Zhang et al (2011) proposed extensions allowing for left-truncated data. Estimation of the regression

coefficients in a proportional subdistribution hazards model using a multiple imputation approach was investigated by Ruan and Gray (2008). The multiple imputation approach was also performed drawing additional bootstrap variables to account for the uncertainty in the Kaplan-Meier estimate of the censoring time distribution. Both approaches revealed similar results, which were also comparable to results obtained from the IPCW approach.

4.2.4 Available software

A Fine and Gray regression model assuming proportional subdistribution hazards can be estimated in the statistical software packages R and SAS . The Fine and Gray regression is included in the R library *cmprsk* provided by Gray (2010) and can be performed using the function *crr*, or in the library *timereg* (Scheike and Martinussen, 2006; Scheike and Zhang, 2011) as a special case in the function *comp.risk*. A macro called *%PSHREG* for proportional and non-proportional subdistribution hazards regression is available in SAS. The multiple imputation approach by Ruan and Gray (2008), mentioned in Section 4.2.3, is implemented in the R library *kmi* (Allignol, 2011).

4.3 Differences between cause-specific and subdistribution hazards regression

The two hazard-based regression approaches, the cause-specific and the subdistribution hazards regression, are the most popular methods for analysis of competing risks data in medical settings. Due to the similarity of the approaches, the regression coefficients obtained from the regression models are often interpreted in an equal manner without considering that the methods focus on different quantities, namely either the cause-specific or the subdistribution hazard. Depending on the amount of competing events and on the covariate effects on the competing events, the two approaches might provide substantially different regression coefficients, as the cause-specific hazards regression aims on the instantaneous risk, whereas the subdistribution hazard is directly linked to the cumulative incidence function. These differences are displayed and discussed for some simulated examples. Other illustrations can be found in Putter et al (2007), Allignol et al (2011), or Dignam et al (2012).

For each scenario competing risks data with two possible endpoints, one event of interest (k=1) and one competing event (k=2), with cause-specific hazards depending on one binary covariate with groups called A (X=0) and B (X=1) were generated for 10,000 subjects. Time-constant cause-specific hazard rates were defined for both groups, so the assumption of proportionality holds for the cause-specific hazards, leading to time-independent cause-specific hazard ratios. For convenience, only administrative censoring after five years was considered. Numbers of patients at risk are displayed under the corresponding figures for both groups to illustrate the influence of competing events on the risk set.

The cause-specific hazard ratio and the subdistribution hazard ratio will generally be different and proportionality for one of these measures contradicts proportionality for the other one. For analysis of the simulated data, proportional hazards regression models for the cause-specific and the subdistribution hazards as described in Section 4.1 and 4.2 were applied, although the assumption of proportionality is violated for the subdistribution

hazards model. The estimated subdistribution hazard ratio can be interpreted as average subdistribution hazard ratio as discussed before (see Latouche et al, 2007; Grambauer et al, 2010; Hjort, 1992).

Scenario 1
In the first scenario the standard survival model with one possible endpoint is shown, i.e. $\lambda_2(t|X{=}0){=}0$ for both groups. Cause-specific hazard rates for $k{=}1$, which is the only possible type of failure in that case, were chosen to be $\lambda_1(t|X{=}0) = 0.2$ and $\lambda_1(t|X{=}1) = 0.4$, translating to a hazard ratio of $\exp(\beta_1) = HR_{k=1} = 2$. As no competing risks are present, cause-specific and subdistribution hazards and consequently cause-specific and subdsitribution hazard ratios are equal in this case. The cumulative incidence functions for the event of interest, displayed in Figure 4.1 (a), are monotonously linked to the corresponding hazard functions as described in Section 2. In the simulated data example with 10,000 subjects (5,000 per group) a hazard ratio of $\exp(\hat{\beta}_1) = 2.01$ was estimated.

Scenario 2
In a second example the difference between the regression coefficients estimated from a proportional cause-specific hazards and a proportional subdistribution hazards regression model in a scenario with two possible endpoints, but a group difference only for the event of interest, was investigated. The cause-specific hazards for the event of interest were chosen to be $\lambda_1(t|X{=}0) = 0.2$ and $\lambda_1(t|X{=}1) = 0.4$, so a cause-specific hazard ratio of 2 was expected for the event of interest. For the competing event ($k{=}2$) the hazard rates were chosen to be equal for both groups $\lambda_2(t|X{=}0) = \lambda_2(t|X{=}1) = 0.3$, implying no group effect on the risk for the competing event. As was to be expected, the estimated cause-specific hazard ratio for the event of interest was close to 2, namely $\exp(\hat{\beta}_1) = 2.01$, the estimated subdistribution hazard ratio for the event of interest was $\exp(\hat{\eta}_1) = 1.81$, which is slightly smaller than the estimated cause-specific hazard ratio due to the different risk sets used. The estimated cumulative incidence functions for both groups are shown in Figure 4.1 (b). The values of the estimated cumulative incidence functions are smaller than in the first example, as less events of interest were observed due to the presence of competing events, which can also be seen by the smaller numbers of individuals under risk at given timepoints. As the cause-specific hazard for the competing event is the same for both groups, the estimated cumulative incidence functions do not cross.

Scenario 3
In the third scenario the cause-specific hazards for the event of interest ($k{=}1$) for both groups were set as in the previous scenarios, giving a cause-specific hazard ratio for the event of interest of $\exp(\beta_1) = HR^{cs}_{k=1} = 2$. The hazard ratio for the competing event ($k{=}2$) was defined to be even larger with the cause-specific hazard in group B being 0.8 and the hazard for group A being 0.2, translating a cause-specific hazard ratio for the competing event of $\exp(\beta_2) = HR^{cs}_{k=2} = 4$. That scenario corresponds to an illustration presented by Putter et al (2007). The cumulative incidence functions for event $k{=}1$ are displayed in Figure 4.1 (c). Due to the higher amount of competing events in group B ($X{=}1$) compared to group A ($X{=}0$), the number of patients at risk is decreasing more slowly in group A. Therefore, a higher incidence of events of interest was observed in group A, although

patients of group B had a higher cause-specific hazard for experiencing an event of type k=1. In that situation the higher cause-specific hazard of group B compared to group A does not translate into a higher incidence of events of type 1 in group B for late timepoints. Analysis of the simulated data gave an estimated cause-specific hazard ratio of 1.99, but a subdistribution hazard ratio of 0.82, revealing different signs of the regression coefficients. The covariate effect on the subdistribution hazard has to be interpreted as time-averaged effect, since the assumption of proportional subdistribution hazards is violated. In the subdistribution hazards regression model, regression coefficients are directly linked to the cumulative incidence function. Since the subdistribution hazard for the event of interest is higher for group A than for group B for most timepoints, a higher average subdistribution hazard for group A is estimated, translating to an average subdistribution hazard ratio smaller than one. Cause-specific hazards regression, representing the covariate effect on the instantaneous risks, and subdistribution hazards regression, showing the effect on the cumulative incidence function, lead to different conclusions regarding the covariate effect on the event of interest.

Scenario 4

In a fourth scenario the setting was chosen similar to Scenario 3, but with a much lower cause-specific baseline hazard for the competing event ($\lambda_2(t|X{=}0) = 0.05$, $\lambda_2(t|X{=}1) = 0.2$), leading to a smaller amount of observed events of type k=2. In the simulations 6029 events of interest were observed (2870 in group X=0, 3159 in group X=1), but only 2297 individuals failed from a competing event (704 in group X=0, 1593 in group X=1). As was to be expected for that case, the difference between estimated cause-specific and subdistribution hazard ratios was smaller than in Scenario 3 with $\exp(\hat{\beta}_1) = 1.95$ and $\exp(\hat{\eta}_1) = 1.28$. The estimated cumulative incidence functions obtained from the simulated dataset are shown in Figure 4.1 (d).

Scenario 5

The cause-specific hazard rates for the event of interest were chosen to be equal for both groups, leading to a cause-specific hazard ratio of one ($\lambda_1(t|X{=}0) = 0.4$, $\lambda_1(t|X{=}1) = 0.4$, $\exp(\beta_1) = HR^{cs}_{k=1} = 1$). For the competing event a cause-specific hazard ratio of $\exp(\beta_2) = 3$ was chosen for the simulation ($\lambda_2(t|X{=}0) = 0.1$, $\lambda_2(t|X{=}1) = 0.3$). The corresponding cumulative incidence functions are displayed in Figure 4.1 (e). Due to the different risks for the competing event, leading to a higher number of competing events in group B than in group A, the number of patients at risk decreases faster in group B. Therefore, a higher incidence of events of interest was observed in group A compared to group B. A cause-specific hazard ratio of 1.03 was estimated, whereas subdistribution hazards regression revealed a hazard ratio of 0.70, since the cumulative incidence curves differ between both groups. In such a situation careless interpretation of the subdistribution hazards regression coefficient might lead to biological implausible conclusions, interpretation of the cause-specific hazards regression coefficient for the event of interest, ignoring the effect on the competing event, will miss important information on the group difference regarding the other type of event and consequently on the event probabilities for the event of interest.

Scenario 6
The cause-specific hazard ratio for both types of event were chosen to be of opposite direction with ($\lambda_1(t|X{=}0) = 0.2$, $\lambda_1(t|X{=}1) = 0.4$, $\exp(\beta_1) = HR^{cs}_{k=1} = 2$) and ($\lambda_2(t|X{=}0) = 0.4$, $\lambda_2(t|X{=}1) = 0.2$, $\exp(\beta_2) = HR^{cs}_{k=2} = 0.5$). As this leads to a difference in the estimated cumulative incidence functions for $k{=}1$, which are shown in Figure 4.1 (f), that is larger than in the absence of competing events, the estimated subdistribution hazard ratio is larger than the estimated cause-specific hazard ratio with $\exp(\hat{\eta}_1) = 2.64$ and $\exp(\hat{\beta}_1) = 2.06$. Due to the opposite direction of the cause-specific hazard ratios for both event types, leading to the same overall hazard, which is defined as the sum of the cause-specific hazards for both types of event as described in Equation 3.7, the number of patients at risk, denoted under the figure, are similar in both groups for all considered timepoints.

Summary and discussion of simulation results
The simulations described above revealed, that substantial differences in the results of cause-specific and subdistribution hazards regression may be present in certain scenarios. Careless interpretation of the estimated regression coefficients may lead to wrong conclusions regarding associations between covariates and risks or event probabilities. Therefore, investigators should be aware of differences between cause-specific hazards and subdistribution hazards regression to avoid misuse of the methods and misinterpretation of obtained results. Beyersmann et al (2007) applied both regression models to a real data example for investigation of occurrence of blood stream infection during neutropenia after peripheral blood stem-cell transplantation, and compared and discussed differences in the methods and in the obtained results. Latouche et al (2013) recommended to present covariate effects obtained from cause-specific hazards regression models for all possible types of event and from a subdistribution hazards regression model for the event of interest, accompanied by estimates of the cumulative incidence functions, to assess whether there is a direct effect of the covariate of interest on the cumulative incidence function (as e.g. in Scenario 2) or an indirect effect caused by an effect on the competing event(s) (as in Scenario 5). Presentation of results obtained from the different regression models and display of the cumulative incidence functions should avoid pitfalls and possible misinterpretations discussed in the examples above.

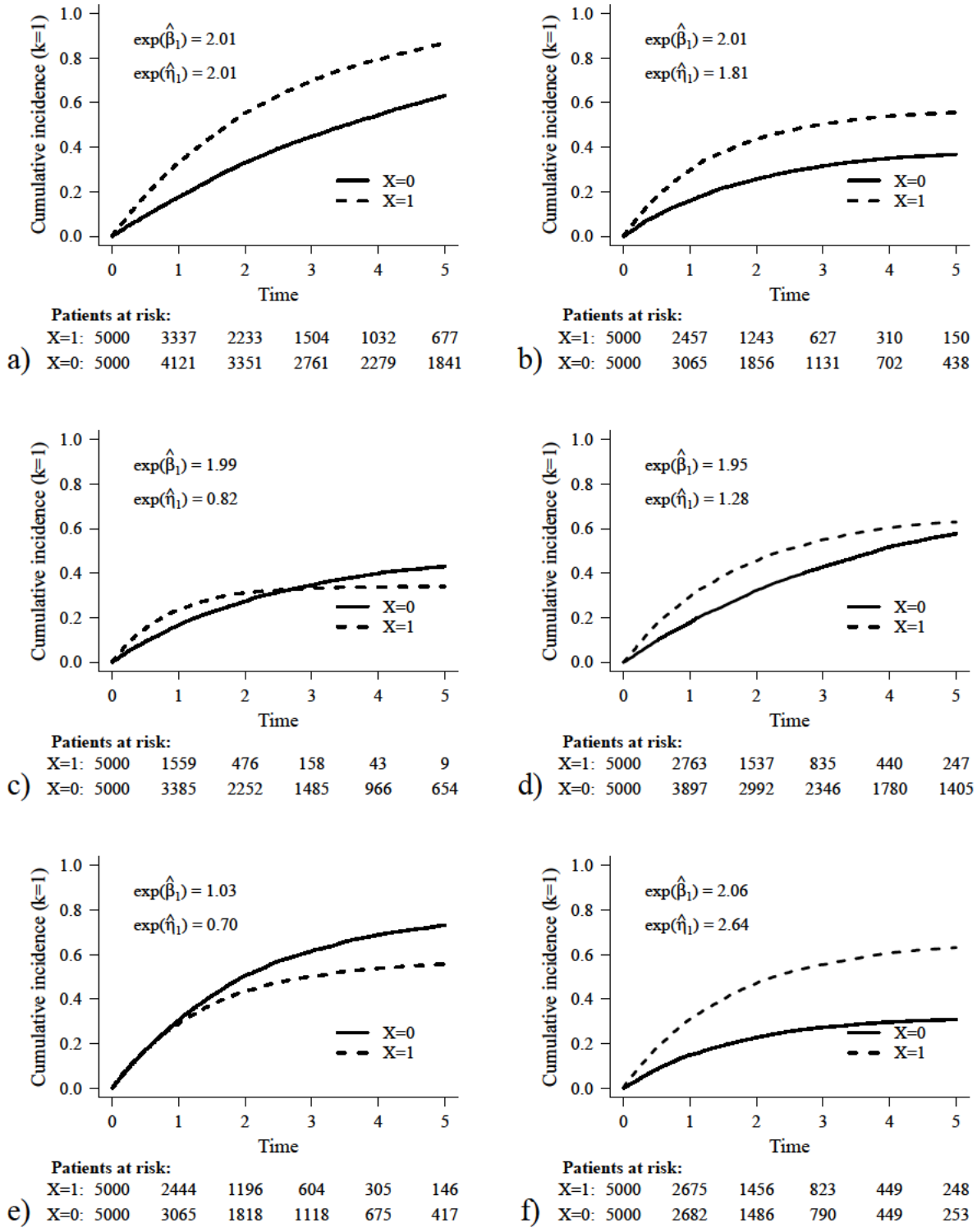

Figure 4.1: Cumulative incidence functions for the event of interest (k=1) of simulated data illustrating the differences in results obtained from cause-specific and subdistribution hazards regression in different scenarios. Estimated cause-specific and subdistribution hazard ratios are presented for all scenarios. Number of individuals at risk in both groups are displayed below the figures.

4.4 The mixture model approach

4.4.1 Background and notation

An alternative approach for the analysis of time-to-event data in the presence of competing risks was introduced by Larson and Dinse in 1985. They proposed to factorize the joint distribution of event time and type of event, that cannot be estimated from observable data, into the marginal event type distribution and the conditional distribution of event times given the type of event

$$P(D,T) = P(D)\,P(T|D), \tag{4.13}$$

where D is a random variable for the type of event and T a random variable for the event time. In this work quantities of the conditional event time distributions will be denoted as follows:

- $\overline{f}_k(t) = f(t|D{=}k)$ is the density function of the conditional event time distribution given an individual failed from an event of type k.
- $\overline{F}_k(t) = F(t|D{=}k)$ is the cumulative density function for the conditional event time distribution given an event of type k, which is a proper distribution function with $\lim_{t\to 0} \overline{F}_k(t) = 0$ and $\lim_{t\to\infty} \overline{F}_k(t) = 1$ for all event types $k = \{1,\ldots,K\}$.
- $\overline{S}_k(t) = S(t|D{=}k)$ is the survivor function of the conditional event time distribution given an event of type k, equalling $1{-}\overline{F}_k(t)$.
- $\overline{h}_k(t) = h(t|D{=}k)$ is the hazard function of the conditional event time distribution.
- $\overline{H}_k(t) = H(t|D{=}k) = \int_0^t h(s|D{=}k)\mathrm{d}s$ is the cumulative hazard function of the conditional event time distribution.

The marginal event type probability $P(D{=}k)$ will be denoted as π_k. In the case of two possible types of event $\pi_2 = 1{-}\pi_1$.
The cumulative incidence function for event type k can be derived from a mixture model as

$$\underline{F}_k(t) = \pi_k\,\overline{F}_k(t). \tag{4.14}$$

The overall survivor function, representing an individual's probability of being free from any event up to time t, is

$$S_{ov.}(t) = \sum_{k=1}^{K} \pi_k\,\overline{S}_k(t). \tag{4.15}$$

For estimation of a mixture model, assumptions for the conditional event time distributions have to be made. Different parametric and semi-parametric approaches were proposed in the literature, some of which are presented and discussed in Sections 4.4.2 and 4.4.3. Generally, parameter estimation in the mixture model approach is performed by numerical maximization of the likelihood or the log-likelihood function. The contribution of individual

i to the mixture model likelihood in the case of K possible types of event and without explicit notation of covariate effects can be denoted as

$$\begin{aligned} L_i = & \left(\pi_1 \, \overline{f}_1(t_i)\right)^{I(d_i=1)} \\ & \times \left(\pi_2 \, \overline{f}_2(t_i)\right)^{I(d_i=2)} \\ & \times \ldots \\ & \times \left(\pi_K \, \overline{f}_K(t_i)\right)^{I(d_i=K)} \\ & \times \left(\pi_1 \, \overline{S}_1(t_i) + \pi_2 \, \overline{S}_2(t_i) + \ldots + \pi_K \, \overline{S}_K(t_i)\right)^{I(d_i=0)} = \\ = & \prod_{k=1}^{K} \left(\pi_k \, \overline{f}_k(t_i)\right)^{I(d_i=k)} \left(\sum_{k=1}^{K} \pi_k \, \overline{S}_k(t_i)\right)^{I(d_i=0)} , \end{aligned} \tag{4.16}$$

with d_i=0 indicating a censored observation. Usually parameters are estimated by maximization of the log-likelihood

$$\begin{aligned} ll = & \ln(L) = \\ = & \sum_{i=1}^{n} \left[\sum_{k=1}^{K} I(d_i=k) \Big(\ln(\pi_k) + \ln\big(\overline{f}_k(t_i)\big) \Big) + I(d_i=0) \, \ln\left(\sum_{k=1}^{K} \pi_k \, \overline{S}_k(t_i) \right) \right] , \end{aligned} \tag{4.17}$$

applying a Newton-Raphson-type or an expectation-maximization (EM) algorithm.
The influence of covariates on the event type probabilities and on the conditional event time distributions can be incorporated by regressing on parameters of the event type distribution, e.g. using a multinomial logistic regression model (see e.g. Fahrmeir and Tutz, 2001), and on parameters of the conditional event time distributions, which is further discussed in the following sections. While the set of covariates assumed to influence event type probabilities and conditional event time distributions, given the type of event, do not have to be same, the set of covariates is assumed not to vary here, so no additional index for the vector of covariates $\mathbf{x}$ will be used.
Different parametric mixture models will be presented in Section 4.4.2, namely a mixture model using piecewise exponential distributions for the conditional event times as originally proposed by Larson and Dinse and mixture models assuming the conditional event times to follow exponential, Weibull, or generalized gamma distributions, which were introduced for standard survival analysis in Section 2.4.
Lau et al (2011) presented, how cause-specific and subdistribution hazards and consequently hazard ratios can be estimated from a mixture model. This will be described in Section 4.5. A new proposal for a mixture model using penalized spline functions, allowing flexible estimation of conditional hazard rates, will be presented in Section 5. The newly proposed approach was compared to the parametric mixture models assuming conditional event times to follow Weibull or generalized gamma distributions in a simulation study, regarding the estimation of cause-specific and subdistribution hazards with a special focus on numerical stability of the estimating procedures. The simulation study and the results are described and discussed in Section 7.

4.4.2 Parametric mixture models

Piecewise exponential mixture model

In their original article published in 1985, Larson and Dinse considered a piecewise exponential distribution for the conditional event times. In a piecewise exponential model, which was described for the common survival setting by Friedman (1982), the hazard rate is assessed for M different time intervals and is assumed to be constant in each interval

$$\overline{h}_k(t) = \exp(\alpha_{k,m}) \quad \text{for all } t \text{ in interval } m,\ m = \{1, \ldots, M\}, \tag{4.18}$$

with $\alpha_{k,m}$ representing the log-hazard rate in time interval m of the conditional event time distribution for a given event of type k.
To incorporate effects of covariates on the event type probabilities and the event time distributions for a given type of event, Larson and Dinse proposed to use a multinomial logistic regression model for the marginal event type distribution and to regress on the piecewise constant hazard rates. The marginal probability for an event of type k, depending on a set of covariates $\mathbf{x}$ including an intercept, can be denoted as

$$P(D{=}k|\mathbf{x}) = \pi_k(\mathbf{x}) = \frac{\exp(\boldsymbol{\mu_k}^\top \mathbf{x})}{\sum_{l=1}^{K} \exp(\boldsymbol{\mu_l}^\top \mathbf{x})}, \tag{4.19}$$

where $\boldsymbol{\mu_K}$ is set to be a vector of zeros to avoid redundancies.
In the case of two possible types of event ($K{=}2$), the probability for an event of type $k{=}1$ can be expressed as

$$P(D{=}1|\mathbf{x}) = \pi_1(\mathbf{x}) = \frac{\exp(\boldsymbol{\mu}^\top \mathbf{x})}{1 + \exp(\boldsymbol{\mu}^\top \mathbf{x})} \tag{4.20}$$

and the probability for an event of type $k{=}2$ as

$$P(D{=}2|\mathbf{x}) = \pi_2(\mathbf{x}) = \frac{1}{1 + \exp(\boldsymbol{\mu}^\top \mathbf{x})}. \tag{4.21}$$

Covariate effects on the conditional event time distributions can be considered by modelling the log-hazard rates, e.g. via $\alpha_{k,m} = \boldsymbol{\beta}_{\boldsymbol{k,m}}^\top \mathbf{x}$.
Larson and Dinse presented an EM algorithm for joint estimation of the regression coefficients from the logistic regression model and the piecewise exponential regression models. Although the piecewise exponential model appears attractive due to its flexibility and ease of interpretation, its use is often questioned due to the biologically implausible jumps of the hazard function at certain timepoints. Moreover, the results of the piecewise exponential approach strongly rely on the number and spacing of the considered time intervals.

Exponential mixture model

When the conditional event times are assumed to follow exponential distributions, the conditional hazard rates $\overline{h}_k(t)$ are assumed to be constant over time. The exponential distribution is described in Section 2.4.1 for the standard survival setting with one possible endpoint and its density and survival function are defined in Equations 2.18 and 2.19. When no covariate effects are investigated, only one parameter, namely an estimate for the

conditional hazard rate or the conditional log-hazard rate, respectively, has to be derived for definition of each conditional event time distribution. The log-likelihood for a mixture model for two possible types of event without consideration of covariates, assuming the conditional event times to follow exponential distributions, can be denoted as

$$\begin{aligned} ll = \sum_{i=1}^{n} \Big[& I(d_i{=}1) \left(\ln(\pi_1) + \ln(\lambda_1) - \lambda_1 t_i \right) + \\ & I(d_i{=}2) \left(\ln(1{-}\pi_1) + \ln(\lambda_2) - \lambda_2 t_i \right) + \\ & I(d_i{=}0) \Big(\ln \left(\pi_1 \, \exp(-\lambda_1 t_i) + (1{-}\pi_1) \, \exp(-\lambda_2 t_i) \right) \Big) \Big]. \end{aligned} \tag{4.22}$$

To assess the influence of covariates of interest, a logistic regression model as defined in Equation 4.19 can be used to model covariate effects on the marginal event type distribution. The covariate influence on the conditional event time distribution can be estimated by modelling the conditional hazard rate

$$\overline{h}_k(t|\mathbf{x}) = \exp\big(\boldsymbol{\beta}_{\boldsymbol{k}}^{\top}\mathbf{x}\big), \tag{4.23}$$

where $\mathbf{x}$ is the vector of covariates with an intercept term and $\boldsymbol{\beta}_{\boldsymbol{k}}$ is the vector of regression coefficients, representing the conditional log-hazard ratio for a given event of type k. When P covariates are considered, $(K{-}1) \times (P{+}1)$ coefficients have to be estimated for the marginal event type distribution and $K \times (P{+}1)$ coefficients for the conditional event time distributions resulting in a total number of $(2K{-}1) \times (P{+}1)$ parameters.

Weibull mixture model

The Weibull distribution, as defined in Section 2.4.2 with density and survivor functions as presented in Equations 2.21 and 2.22, is a two-parameter event time distribution allowing more flexible hazard functions than the one-parameter exponential distribution. In the absence of covariates the log-likelihood for a mixture model with conditional event times following Weibull distributions can be written for two possible types of event as

$$\begin{aligned} ll = \sum_{i=1}^{n} \Big[& I(d_i{=}1) \Big(\ln(\pi_1) + \ln(\lambda_1) + \ln(\alpha_1) + (\alpha_1{-}1)\big[\ln(\lambda_1) + \ln(\alpha_1) \big] - (\lambda_1 t_i)^{\alpha_1} \Big) + \\ & I(d_i{=}2) \Big(\ln(1{-}\pi_1) + \ln(\lambda_2) + \ln(\alpha_2) + (\alpha_2{-}1)\big[\ln(\lambda_2) + \ln(\alpha_2) \big] - (\lambda_2 t_i)^{\alpha_2} \Big) + \\ & I(d_i{=}0) \Big(\ln\big(\pi_1 \, \exp(-(\lambda_1 t_i)^{\alpha_1}) + (1{-}\pi_1) \, \exp(-(\lambda_2 t_i)^{\alpha_2})\big) \Big) \Big]. \end{aligned} \tag{4.24}$$

Covariate influence on the event type probabilities will be assessed as described before. In a Weibull regression model the influence of covariates on the parameters λ_k can be assessed as in the exponential model, with the shape parameters α_k not depending on covariate values. Additionally, the covariate effect on parameters α_k can be assessed, allowing more flexible conditional event time distributions, but relying on more parameters to be estimated. In a Weibull mixture model, regressing on the marginal event type distribution and on the

parameters λ_k of the conditional event time distributions, $(K-1)\times(P+1) + K\times(P+1) + K = (2K-1)\times(P+1) + K$ parameters have to be estimated. If the covariate influence on the shape parameters α_k is also modelled, the number of parameters to be estimated increases to $(K-1)\times(P+1) + K\times(P+1) + K\times(P+1) = (3K-1)\times(P+1)$.

Generalized gamma mixture model

In their article on the estimation of cause-specific and subdistribution hazard rates and hazard ratios from a mixture model, Lau et al (2011) propose to use a flexible parametric survival distribution as the three-parameter generalized gamma distribution. In this work the parametrization of the generalized gamma distribution that is considered by Lau et al and was investigated and discussed by Cox et al (2007) is used. The density function and the survivor function are shown in Section 2.4.3 in Equations 2.24 and 2.25. The log-likelihood function to be maximized in order to obtain parameter estimates for a mixture model, assuming the conditional event times to follow generalized gamma distributions, is

$$\begin{aligned} ll = \sum_{i=1}^{n} \Big[& I(d_i{=}1) \Big(\ln(\pi_1) + \ln(|\nu_1|) - \ln(\tilde{\alpha}_1) - \ln(t_i) - \ln\big(\Gamma(\nu_1^{-2})\big) + \\ & \qquad + \nu_1^{-2}\big[\ln(\nu_1^{-2}) + \frac{\nu_1}{\tilde{\alpha}_1}\big(\ln\lambda_1 + \ln(t_i)\big) - \nu_1^{-2}(\lambda_1\, t_i)^{\nu_1/\tilde{\alpha}_1}\big]\Big) + \\ & + (I(d_i{=}2) \Big(\ln(1{-}\pi_1) + \ln(|\nu_2|) - \ln(\tilde{\alpha}_2) - \ln(t_i) - \ln\big(\Gamma(\nu_2^{-2})\big) + \\ & \qquad + \nu_2^{-2}\big[\ln(\nu_2^{-2}) + \frac{\nu_2}{\tilde{\alpha}_2}\big(\ln\lambda_2 + \ln(t_i)\big) - \nu_2^{-2}(\lambda_2\, t_i)^{\nu_2/\tilde{\alpha}_2}\big]\Big) + \\ & + I(d_i{=}0)\, \ln\big(\pi_1\, \overline{S}_1(t_i) + (1{-}\pi_1)\, \overline{S}_2(t_i)\big)\Big], \end{aligned} \tag{4.25}$$

where $\overline{S}_k(t)$ is the conditional survivor function for a given event of type k as defined in Equation 2.25.
Covariate effects on the conditional event time distributions were assessed in Lau et al (2011) by modelling the location parameters λ_k via

$$\lambda_k = \exp(-\boldsymbol{\beta_{k,\lambda}}^{\top}\mathbf{x}). \tag{4.26}$$

The shape and scale parameters were assumed to be independent of the covariate values. In Cox et al (2007) a generalized gamma regression model was used in a standard survival setting with one possible type of event assessing covariate effects on all three parameters, which was called "saturated generalized gamma model" by Cox et al. In the competing risks mixture model $(K-1)\times(P+1)+K\times(P+1)+2K = (2K-1)\times(P+1)+2K$ parameters have to be estimated, when only the location parameters λ_k are allowed to depend on covariates. In a saturated generalized gamma mixture model, allowing all three parameters of the conditional event time distributions to depend on covariates, the number of parameters to be estimated increases to $(K-1)\times(P+1) + 3\times K\times(P+1) = (4K-1)\times(P+1)$.
Both generalized gamma mixture models, the model assessing covariate effects on the location parameters λ_k only and the saturated model, were investigated in the simulation study presented in Section 7.

4.4.3 Semi-parametric mixture models

In order to allow more flexible modelling of the conditional baseline hazard functions, different semi-parametric approaches, assuming proportional conditional hazard rates, were proposed in the literature. The conditional hazard rate for a given event type k can be denoted as known from common Cox regression as

$$\overline{h}_k(t) = \overline{h}_{k,0}(t)\ \exp(\boldsymbol{\beta_k}^\top \mathbf{x}), \tag{4.27}$$

with $\overline{h}_{k,0}(t)$ being the conditional baseline hazard function.
One early approach was proposed by Kuk and Chen (1992), who approximated the marginal likelihood by Monte Carlo methods drawing random variables for censored observations. The approach was criticised later as the results were heavily depending on the sampling mechanism (see e.g. Escarela and Bowater, 2008).
Ng and McLachlan (2003) proposed a semi-parametric proportional hazards mixture model allowing maximum likelihood estimation of the parameters from the full likelihood by using an expectation conditional maximization (ECM) algorithm, treating the conditional baseline hazard functions as high-dimensional nuisance parameters. In the ECM algorithm the expectation of the complete data log-likelihood is estimated in the E-step, considering the probabilities of failing from the given event types for the censored individuals conditional on the observed data and the current parameter estimates. In the M-step the expected log-likelihood considering the completely observed data and the derived expectations for censored observations is maximized in order to obtain updated estimates. The E-step and the M-step are altered until some predefined convergence criterion is reached. It can be shown that the value of the likelihood calculated at the current maximum likelihood estimate is increased for each step. In order to assess variance estimates and confidence intervals, Ng and McLachlan recommended to consider bootstrap samples by drawing subsamples from individuals that failed from the different event types and from censored observations with the number of samples equalling the number of observed events and censored observations, respectively. In a simulation study provided in their article the proposed algorithm was superior to parametric models, if the true underlying baseline hazard was non-monotonous, but they did not investigate very flexible parametric mixture models as the generalized gamma mixture model. Although the baseline hazards of the conditional event time distributions are not needed for estimation of the regression coefficients in the proportional hazards approach, they have to be estimated in order to derive estimates for the cumulative incidence functions from the mixture model.
A similar approach was published by Escarela and Bowater (2008) using an extension of the EM algorithm provided by Larson and Dinse (1985). The large sample properties of the approaches by Ng and McLachlan (2003) and Escarela and Bowater (2008) were further investigated by Hernandez-Quintero et al (2011), establishing asymptotic normality of the derived estimates and proposing a consistent variance estimator.
In this work the approach by Ng and McLachlan (2003) was applied to the data example presented in Section 8. Regression coefficients were estimated using the ECM algorithm and bootstrap samples were drawn in order to obtain confidence intervals. The R code used for the analysis is sketched in Section C.3 of the Appendix.

4.5 Estimating cause-specific and subdistribution hazard rates from a mixture model

The most frequently used regression approaches for the analysis of competing risks data are the hazard based regression models, namely the cause-specific and the subdistribution hazards regression models presented in Section 4.1 and 4.2, and the mixture model approach as presented above. Lau et al (2011) described, how cause-specific and subdistribution hazards and consequently hazard ratios can be derived from a mixture model. The procedure will be presented in this section. Lau et al (2011) proposed to use a flexible parametric event time distribution, like the generalized gamma distribution, to model the conditional hazard rates in order to allow for various shapes of the conditional hazard functions and consequently of the estimated cause-specific and subdistribution hazard rates and ratios. In Section 5 an alternative approach using penalized B-spline functions (P-splines) for flexible estimation of conditional hazard rates will be presented.

4.5.1 Estimating the cause-specific hazard rate

The cause-specific hazard rate for event type k, denoted as $\lambda_k(t|\mathbf{x})$, can be estimated as the quotient of the subdensity function $\underline{f}_k(t|\mathbf{x})$ and the overall survivor function $S_{ov.}(t|\mathbf{x})$, which represents an individual's probability of being free from any event up to time t. The subdensity function for event type k, which is defined through $\underline{f}_k(t|\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}t}\underline{F}_k(t|\mathbf{x})$, can be expressed using quantities derived from a mixture model, namely as the product of the density function of the conditional event time distribution for a given event of type k, and the probability for an event of type k

$$\underline{f}_k(t|\mathbf{x}) = \overline{f}_k(t|\mathbf{x})\, P(D{=}k|\mathbf{x}) = \overline{f}_k(t|\mathbf{x})\, \pi_k(\mathbf{x}). \tag{4.28}$$

Estimates for $\overline{f}_k(t|\mathbf{x})$ and $P(D{=}k|\mathbf{x})$ can be derived from the parameters estimated by maximizing the log-likelihood of the mixture model as presented in Equation 4.16. The overall survivor function $S_{ov.}(t|\mathbf{x})$ is a weighted average of the conditional survivor functions

$$S_{ov.}(t|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x})\, \overline{S}_k(t|\mathbf{x}) \tag{4.29}$$

and can be estimated accordingly from coefficients estimated from a mixture model.
The cause-specific hazard function for event type k, given the vector of covariates $\mathbf{x}$, is the quotient of the subdensity function and the overall survivor function

$$\lambda_k(t|\mathbf{x}) = \frac{\underline{f}_k(t|\mathbf{x})}{S_{ov.}(t|\mathbf{x})}. \tag{4.30}$$

Cause-specific hazard ratios can now be determined as a function of time by inserting the covariate values of interest. To derive the cause-specific hazard ratio for event type k for a group variable X_1 with two possible outcomes ($X_1{\in}\{0,1\}$), adjusted for other covariates $X_2, \ldots, X_P$, the quotient of the estimated cause-specific hazards, inserting either zero or

one for X_1 and the mean of the whole study population for all other $P-1$ covariates, is calculated

$$\widehat{HR}_k^{cs}(t) = \frac{\hat{\lambda}_k\big(t|\mathbf{X}=(1,\bar{x}_2,\ldots,\bar{x}_P)^\top\big)}{\hat{\lambda}_k\big(t|\mathbf{X}=(0,\bar{x}_2,\ldots,\bar{x}_P)^\top\big)}. \tag{4.31}$$

4.5.2 Estimating the subdistribution hazard rate

Subdistribution hazard rates for event type k, given a vector of covariates $\mathbf{x}$, and consequently subdistribution hazard ratios can be estimated in a similar manner as the cause-specific hazard ratios. The subdistribution hazard rates can also be derived using quantities of a mixture model, by dividing the subdensity function $\underline{f}_k(t|\mathbf{x})$ through one minus the cumulative incidence function

$$\gamma_k(t|\mathbf{x}) = \frac{\underline{f}_k(t|\mathbf{x})}{1-\underline{F}_k(t|\mathbf{x})}. \tag{4.32}$$

The cumulative incidence function can be obtained from a mixture model as shown in Equation 4.14. A subdistribution hazard ratio can be derived analogously to the cause-specific hazard ratio (see Equation 4.31) by diving the estimated subdistribution hazard rates for corresponding covariate values

$$\widehat{HR}_k^{sd}(t) = \frac{\hat{\gamma}_k\big(t|\mathbf{X}=(1,\bar{x}_2,\ldots,\bar{x}_P)^\top\big)}{\hat{\gamma}_k\big(t|\mathbf{X}=(0,\bar{x}_2,\ldots,\bar{x}_P)^\top\big)}. \tag{4.33}$$

4.6 Vertical modelling

In 2010 another approach factorizing the joint distribution of event times and event types into a marginal and a conditional distribution was introduced by Nicolaie et al. In contrast to the mixture model approach presented above, in the so called vertical modelling approach the joint distribution of event times and types is represented as the product of the marginal event time distribution and the conditional event type distribution

$$P(T,D) = P(T)\,P(D|T). \tag{4.34}$$

Here again T is a random variable for the event time and D a random variable for the type of event.
In the vertical modelling approach the marginal event time distribution can either be estimated using a standard survival function estimator as the Kaplan-Meier method or the marginal event time distribution can be estimated considering covariate effects on the overall hazard rate using e.g. a Cox regression model or a parametric regression model for time-to-event data, treating events of any type as failures.
The conditional event type distribution provides the so called relative hazards, denoted as $\tilde{\pi}_k(t)$, i.e. the probabilities for events of type $k=\{1,...,K\}$ given any event was observed at time t

$$\tilde{\pi}_k(t) = P(D{=}k\,|\,T{=}t). \tag{4.35}$$

The natural estimate for the relative hazard of event type k for an observed event time $\tilde{t}_i$, denoted as $\hat{\tilde{\pi}}_k(\tilde{t}_i)$, would be the number of observed events of type k at time $\tilde{t}_i$ divided by the total number of events at time $\tilde{t}_i$

$$\hat{\tilde{\pi}}_k(\tilde{t}_i) = \frac{\frac{d_{k\tilde{t}_i}}{R_{\tilde{t}_i}}}{\frac{d_{\tilde{t}_i}}{R_{\tilde{t}_i}}} = \frac{d_{k\tilde{t}_i}}{d_{\tilde{t}_i}} \,, \tag{4.36}$$

with $d_{k\tilde{t}_i}$ denoting the number of observed type k events at time $\tilde{t}_i$, $d_{\tilde{t}_i}$ the total number of events at $\tilde{t}_i$, and $R_{\tilde{t}_i}$ the number of patients at risk at $\tilde{t}_i$.
If events are observed in continuous time, this will lead to a series of zeros and ones and therefore this method will not produce appropriate estimates for the relative hazards in general. In order to obtain adequate and interpretable relative hazard estimates, Nicolaie et al proposed to fit a multinomial (or in the case of two possible endpoints a binomial) logistic regression model to the data, considering time and probably further variables of interest as covariates. In order to allow relative hazards to vary over time, the influence of time can be modelled flexibly. In the article by Nicolaie et al the use of cubic B-spline functions was proposed to obtain flexible estimates for the relative hazards. In the case of K possible event types, the estimate for the relative hazard of the k^{th} event type can be denoted as

$$\hat{\tilde{\pi}}_k(t|\mathbf{x}) = \frac{\exp\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{k}}^\top \mathbf{B}(t) + \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}^\top \mathbf{x}\right)}{\sum_{l=1}^{K} \exp\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{l}}^\top \mathbf{B}(t) + \hat{\boldsymbol{\beta}}_{\boldsymbol{l}}^\top \mathbf{x}\right)} \tag{4.37}$$

$\mathbf{B}(t)$ denotes the set of B-spline basis functions, which are described in more detail in Section 5, $\boldsymbol{\xi_k}$ is the vector of regression coefficients for the B-spline basis functions for event type k, $\boldsymbol{\beta_k}$ is the corresponding vector of regression coefficients indicating influence of covariates $\mathbf{x}$. The regression coefficients can be derived as known for a logistic regression model (see e.g. Fahrmeir and Tutz, 2001), including only subjects with an observed event. Censored observations can only be considered for estimation of the marginal event time distribution. If the number of observed events is sufficient, interaction effects between the covariates of interest and the B-spline components of time can be incorporated, in order to allow for different patterns of relative hazards for relevant groups or patient characteristics. The estimated relative hazards $\hat{\tilde{\pi}}_1(t), \ldots, \hat{\tilde{\pi}}_K(t)$ sum up to one for every timepoint t.
As interpretation of the relative hazards based on the regression coefficients is very difficult due to the complicated structure of the B-spline components, it is recommended to illustrate the results graphically. A plot of the relative hazards, showing estimated probabilities for all K types of event, given any event occurs at time t, can be used to illustrate the estimated pattern of conditional event probabilities. To avoid overinterpretation of relative hazards in time intervals with a small number of events, the relative hazards should always be displayed in conjunction with estimates for the marginal event time distribution.
Nicolaie et al applied the vertical modelling approach to data of leukemia patients from the European Group for Blood and Marrow Transplantation in order to investigate patterns of adverse events as relapse, graft-versus-host disease, or infections after allogeneic hematopoietic stem cell transplantation. In our work, the vertical modelling approach was used to assess the patterns of cardiac and non-cardiac deaths in patients that survived a myocardial infarction with a special interest in differences between patients that were identified as being of high risk for a myocardial infarction and patients that were identified

as being of low risk. A further description of the analysis and the results can be found in Section 8.

4.7 Regression models based on pseudo-observations

Andersen et al (2003) introduced a method for the estimation of covariate effects on state probabilities in multi-state models using pseudo-observations. Since a competing risks model can be interpreted as a special case of a multi-state model, this approach can be adjusted for the competing risks setting as demonstrated by Klein and Andersen (2005). Generally, the pseudo-observation approach can be considered to estimate effects of covariates on any function of event times $f(T)$, if an unbiased estimator $\hat{\theta}$ exists for

$$\theta = E\big(f(T)\big). \tag{4.38}$$

A summary of different methods for survival analysis based on pseudo-observations is presented by Andersen and Perme (2010). Main idea of the approach is to obtain quantities, that allow application of standard methods for data analysis without consideration of censored observations. The estimated pseudo-observations $\hat{\theta}_i$, $i = \{1, \ldots, n\}$, which are assessed via leave-one-out estimates (see e.g. Miller, 1974) for some measure of interest, can be used for that purpose

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(i)}. \tag{4.39}$$

Here $\hat{\theta}$ is the estimated measure of interest using all n observations and $\hat{\theta}^{(i)}$ indicates the estimated measure of interest derived from all but the i^{th} observation. The pseudo-observations can be estimated for one fixed timepoint τ_0 or for a prespecified number of timepoints $\boldsymbol{\tau} = (\tau_1, ..., \tau_H)$. If multiple timepoints are considered, a $n \times H$-matrix of pseudo-observations is obtained. For regression purposes these pseudo-observations $\hat{\theta}_{ih}$ can be used as dependent variable (Klein and Andersen, 2005) in a generalized linear model

$$g(\theta_{ih}|\mathbf{x_i}) = \alpha_h + \boldsymbol{\beta}^\top \mathbf{x_i}, \tag{4.40}$$

where $g(\cdot)$ is a link function as the logit or the complementary log-log function and $\mathbf{x_i}$ is the vector of covariates of subject i. The influence of the covariates on the pseudo-observations, that translates to an influence of the covariates on the measure of interest $f(T)$, can be estimated using adequate methods for generalized linear models. In the case of multiple timepoints the generalized estimation equation approach (GEE, Liang and Zeger, 1986) was proposed for estimation and inference to account for repeated measures on the same subjects in order to obtain robust and valid standard errors under independent censoring. Klein and Andersen (2005) discuss different assumptions for the working covariance matrix used in the GEE model. Since they did not find any relevant effects of the choice of the working covariance on the estimated regression coefficients and standard errors, they proposed the use of an independent working covariance structure.
For the competing risks setting the relevant measure $f(T)$ is the cumulative incidence function for event type k. So for each individual i a pseudo-observation $\hat{\theta}_{ih}$ is derived for each of the predefined timepoints in $\boldsymbol{\tau}$, using the cumulative incidence function estimated from all subjects and the estimate based on all but the i^{th} individual

$$\hat{\theta}_{ih} = n\underline{\hat{F}}_k(\tau_h) - (n-1)\underline{\hat{F}}_k^{(i)}(\tau_h). \tag{4.41}$$

If censoring is absent in the whole dataset, the pseudo-value indicates, whether subject i failed from cause k up to time τ_h, i.e. $\hat{\theta}_{ih}$=1 if $t_i \leq \tau_h$ and d_i=k or $\hat{\theta}_{ih}$=0, else, and the mean of the pseudo-values for each considered timepoint equals the estimate of the cumulative incidence function. In the presence of censored observations, pseudo-values can be smaller than zero for individuals still under observation, for individuals with a censored observation or for individuals that failed from a competing event, or larger than one after an event of interest was observed, with the actual value depending on the observation time and the amount of censoring. An illustration of pseudo-observations can be found in Andersen and Perme (2010) for different measures of interest, including the cumulative incidence function, and in Section 8.2.5 of this work for the investigated clinical data (Figure 8.3).
When a complementary log-log link is used between the response (the pseudo-values) and the linear predictor, the regression coefficients can be interpreted as subdistribution log-hazard ratios, if all covariates are time-independent (Klein and Andersen, 2005)

$$\ln\big(-\ln(\theta_{ih})\big) = \alpha_h + \boldsymbol{\beta}^\top \mathbf{x_i}. \tag{4.42}$$

The analysis can be performed using the R function *geese* from the R library *geepack* (Højsgaard et al, 2005), that allows to specify a complementary log-log link between response and linear predictor.
SAS and R functions for the computation of pseudo-values for time-to-event data are provided by Klein et al (2008). Recently, Andersen and Perme (2010) recommended not to use pseudo-observations for the analysis of time-to-event data assuming proportional subdistribution hazards, since the Fine and Gray model was identified to be more efficient in a simulation study. However, the pseudo-observation approach can be conducted in more complex situations when standard regression models are not applicable. Furthermore, the use of pseudo-residuals and pseudo-scatterplots for investigation of model assumptions in regression models is encouraged (Perme and Andersen, 2008).
In this work, the pseudo-value approach was applied to data from the clinical cohort study to complement the other regression approaches. Details of the performed analysis and the results as well as a discussion on the pseudo-value approach and a comparison to the other approaches can be found in Section 8.

Chapter 5

A new approach for flexible estimation of cause-specific and subdistribution hazards from a mixture model using penalized spline functions

An alternative approach for estimation of conditional hazard rates in a mixture model is proposed in this section. Instead of assuming the conditional event times to follow a certain parametric distribution, as the exponential, the Weibull or the generalized gamma distribution, or assuming proportional conditional hazard rates without further specification of the conditional baseline hazard rates, the hazard functions of the conditional event time distributions will be estimated using penalized B-spline functions.

5.1 Splines in event time analysis

Spline functions are commonly used for smooth estimation of distribution functions or in regression models to derive relationships between dependent and independent variables without the assumption of linear relationships. The main idea behind the use of spline functions is to use a set of predefined basis functions that are weighted by regression coefficients to approximate the quantity or relationship of interest. A detailed description and discussion on the use of spline functions in statistical models can e.g. be found in Wegman and Wright (1983), Fahrmeir et al (2007) or Hastie et al (2009).
Different applications of spline functions were introduced in the context of event time analysis for smooth estimation of hazard functions. Rosenberg (1995) proposed to use B-spline functions, which are described in detail in Section 5.2.1, to estimate hazard functions in a flexible way. He applied the proposed method to model the hazard of developing AIDS after infection with the HI virus. Kooperberg et al (1995) proposed models for the log-hazard rate using spline functions. They also introduced an algorithm for model selection based on Akaike's information criterion (AIC Akaike, 1974). The presented models were applied to different clinical datasets. Royston and Parmar (2002) described the estimation of a

proportional hazards model and a proportional odds model with the baseline cumulative hazard function or the baseline cumulative odds function estimated flexibly using cubic spline functions. The methods were applied to data of breast cancer patients comparing survival between patient groups with different prognosis and to data from a randomized clinical trial comparing treatments in patients suffering from advanced bladder cancer.
Several applications of spline functions in event time analysis for smooth estimation of covariate effects relaxing the linearity assumption can be found. In the competing risks setting application of spline functions for smooth estimation of covariate effects on cause-specific hazards was presented by Belot et al (2010).

5.2 Modelling hazard functions using cubic B-spline basis functions

In this work cubic B-spline basis functions will be considered for smooth estimation of the conditional hazard rates given the type of event in a competing risks mixture model. For convenience and ease of notation, the approach is presented for a binary covariate X and for two possible types of event $k=\{1,2\}$. Modelling the the hazard functions of the conditional event time distributions in a similar way as described by Rosenberg (1995) for a common survival setting is proposed here and the definition of B-spline basis functions presented there is used.

5.2.1 One possible endpoint

Firstly, the use of B-spline basis functions for hazard estimation in a common survival setting with one possible type of event is considered. With G predefined interior knots, the g^{th} component of the set of B-spline basis functions can be denoted as

$$B_g(t) = (t_{g+4} - t_g) \sum_{h=g}^{g+4} \Big\{ (t_h - t)_+^3 \Big/ \prod_{\substack{i \neq h \\ i=g,\ldots,g+4}} (t_i - t_h) \Big\}, \tag{5.1}$$

where $(z)_+$ returns the value of z, if z is larger than zero and zero, else. For a complete definition of the basis functions, three lower and three upper knots have to be defined. Different recommendations were made on how to set these so called slack knots. Here the distance between the three lower knots was chosen to be defined by the distance between the first observed event time (T_{min}), and the first interior knot, and the distance between the upper knots to be the same as the distance between the last interior knot and the maximum observed event time (T_{max}), as proposed by Fahrmeir et al (2007). The set of cubic B-spline basis functions with five interior knots is illustrated in Figure 5.1 for equidistant knots (a) and for knots depending on estimated quantiles of the event time distribution (b). For each timepoint t in the interval from T_{min} to T_{max} four of the basis functions are larger than zero and for each of these timepoints the unweighted basis functions sum up to one.
In his article Rosenberg proposed to model the hazard function in a standard survival

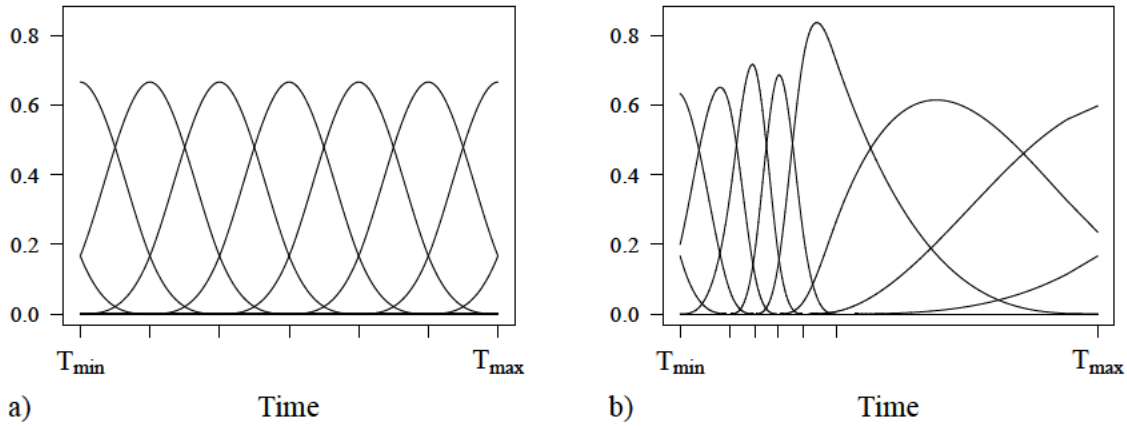

Figure 5.1: Illustration of B-spline basis functions with five interior knots placed equidistantly (left) or placed at observed event time quantiles (right). The ticks on the x-axes indicate the placing of the knots.

setting without consideration of covariate effects as

$$\lambda(t) = \sum_{g=-3}^{G} B_g(t) \exp(\beta_g), \tag{5.2}$$

where $B_g(t)$ is the g^{th} component of the set of B-spline basis functions $\mathbf{B(t)}$ as defined in Equation 5.1, and β_g is a regression coefficient weighting that component. In order to ensure the estimated hazard functions to be positive for all timepoints, the regression coefficients are exponentiated. For a simulated example the true hazard rate and the unweighted B-spline functions are shown in Figure 5.2 (a), the weighted B-spline functions used for approximation of the true underlying hazard and the sum of the components giving an estimate for the hazard function are presented in picture (b) (Figure 5.2 is displayed on page 55).
In order to illustrate a regression model, assessing the influence of covariates on the hazard function, a setting with one binary covariate $X \in \{0, 1\}$ is considered for convenience and ease of notation. The setting is also used for the mixture model approach presented in Section 5.2.2. In this setting the hazard function can be denoted in dependence of the covariate value as

$$\lambda(t|x) = \sum_{g=-3}^{G} B_g(t) \exp(\beta_{0,g}+x\beta_{1,g}), \tag{5.3}$$

so the weights for the g^{th} B-spline component are $\exp(\beta_{0,g})$ for X=0 and $\exp(\beta_{0,g}+\beta_{1,g})$ for X=1. The survivor function, which is needed for maximum likelihood estimation of the regression weights in the presence of censored observations, is

$$S(t|x) = \exp\left(- \sum_{g=-3}^{G} \Big(IB_g(t) - IB_g(T_{min}) \Big) \exp(\beta_{0,g} + x\beta_{1,g}) \right), \tag{5.4}$$

with $IB_g(t)$ as described in Rosenberg (1995)

$$IB_g(t) = -\frac{(t_{g+4} - t_g)}{4} \left(\sum_{h=g}^{g+4} (t_h - t)_+^4 \Big/ \prod_{\substack{i \neq h \\ i=g,\ldots,g+4}} (t_i - t_h) \right). \tag{5.5}$$

The regression coefficients can be estimated by numerically maximizing the log-likelihood (see Equation 2.17)

$$\begin{aligned} ll &= \sum_{i=1}^{n} I(d_i{=}1)\, \ln\big(f(t_i|x_i)\big) + I(d_i{=}0) \ln\big(S(t_i|x_i)\big) = \\ &= \sum_{i=1}^{n} I(d_i{=}1)\, \ln\big(\lambda(t_i|x_i)\big) + \ln\big(S(t_i|x_i)\big) = \\ &= \sum_{i=1}^{n} I(d_i{=}1)\, \ln\left(\sum_{g=-3}^{G} B_g(t_i)\, \exp(\beta_{0,g}{+}x_i\beta_{1,g}) \right) \\ &\qquad\qquad - \sum_{g=-3}^{G} \Big(IB_g(t_i) - IB_g(T_{min}) \Big)\, \exp(\beta_{0,g}{+}x_i\beta_{1,g}). \end{aligned} \tag{5.6}$$

5.2.2 Extension to the mixture model approach

In this section the B-spline approach as described in Section 5.2.1 for a standard survival setting with one possible endpoint is adapted for estimation of conditional hazard rates in a competing risks mixture model, allowing flexible estimation of cause-specific and subdistribution hazards as described in Section 4.5. As a conditional hazard rate for each possible type of event has to be estimated, modelling the hazard rate as presented in Equation 5.3 has to adapted by including an additional index k for the type of event

$$\overline{h}_k(t|x) = \sum_{g=-3}^{G} B_g(t)\, \exp(\beta_{k,0,g} + x\beta_{k,1,g}). \tag{5.7}$$

The relationship between basis functions, covariates and the survivor function presented in Equation 5.4 has to be adapted accordingly

$$\overline{S}_k(t|x) = \exp\left(- \sum_{g=-3}^{G} \Big(IB_g(t) - IB_g(T_{min}) \Big)\, \exp(\beta_{k,0,g} + x\beta_{k,1,g}) \right). \tag{5.8}$$

The vector of regression coefficients in the mixture model, denoted here as $\boldsymbol{\beta_{MM}}$, consists of different components, namely the components for the marginal event type distribution and for the conditional event time distributions weighting the B-spline basis functions, which in the setting with a binary covariate and two possible event types is

$$\boldsymbol{\beta_{MM}} = (\boldsymbol{\mu}, \boldsymbol{\beta_{1,0}}, \boldsymbol{\beta_{1,1}}, \boldsymbol{\beta_{2,0}}, \boldsymbol{\beta_{2,1}})^\top.$$

The log-likelihood to be maximized for parameter estimation in the setting with two possible types of event considering one binary covariate can be written as

$$\begin{aligned} ll = \sum_{i=1}^{n} \Big[& I(d_i{=}1)\Big(\ln\big(\pi(x_i)\big) + \ln\big(\overline{f}_1(t_i|x_i)\big)\Big) + \\ & + I(d_i{=}2)\Big(\ln\big(1{-}\pi(x_i)\big) + \ln\big(\overline{f}_2(t_i|x_i)\big)\Big) + \\ & + I(d_i{=}0)\, \ln\Big(\pi(x_i)\, \overline{S}_1(t_i|x_i) + \big(1{-}\pi(x_i)\big)\, \overline{S}_2(t_i|x_i)\Big)\Big] = \\ = \sum_{i=1}^{n} \Big[& I(d_i{=}1)\Big(\ln\big(\pi(x_i)\big) + \ln\big(\overline{h}_1(t_i|x_i)\big) + \ln\big(\overline{S}_1(t_i|x_i)\big)\Big) + \\ & + I(d_i{=}2)\Big(\ln\big(1{-}\pi(x_i)\big) + \ln\big(\overline{h}_2(t_i|x_i)\big) + \ln\big(\overline{S}_2(t_i|x_i)\big)\Big) + \\ & + I(d_i{=}0)\, \ln\Big(\pi(x_i)\, \overline{S}_1(t_i|x_i) + \big(1{-}\pi(x_i)\big)\, \overline{S}_2(t_i|x_i)\Big)\Big], \end{aligned} \tag{5.9}$$

where $\pi(x)$ denotes the probability for an event of type $k{=}1$ given x, which is modelled using a binary logistic regression model as shown in Equation 4.20. Maximum likelihood estimates for the regression coefficients can be derived by numerical maximization of the log-likelihood function.
The estimated regression coefficients can be used to derive estimates for the conditional hazard functions, the conditional survivor functions and the marginal event type distribution for both groups. Using these quantities, estimates for the subdensity functions, the cumulative incidence functions and the overall survivor functions can be derived as shown in Section 4.5, and consequently the cause-specific and subdistribution hazard rates can be estimated as presented in Equations 4.30 and 4.32.
Choices for the number of basis functions and the spacing of knots, when spline methods are used, are still a matter of research and discussion. A higher number of knots leads to more flexible functions, but also to a higher number of coefficients to be estimated and possibly to overfitting or problems in the numerical maximization procedure. One possibility to reduce these problems is penalizing the roughness of the resulting function by adding a penalty term to the log-likelihood, as it is done in the P-spline approach introduced by Eilers and Marx (1996). This procedure is described in the next section.

5.3 Penalization

In the so called P-spline approach roughness of the estimated hazard functions is controlled by penalizing differences of nearby coefficients. A smoothing parameter μ is introduced for that purpose. Second order differences are considered here to regulate the flexibility of the obtained conditional hazard rates.

5.3.1 One possible endpoint

For estimation of the hazard rate in a standard survival setting using penalized spline functions, the penalty matrix of second order differences $\mathbf{D_2}$ is a $(G{+}4) \times (G{+}2)$ matrix,

with G being the number of interior knots considered for definition of the set of B-spline basis functions, as (G+4) regression coefficients have to be estimated

$$\mathbf{D_2} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{5.10}$$

The parameter μ is used to control the penalization and consequently the smoothness of the estimated hazard function. Parameter estimation can be conducted by maximization of an adapted likelihood or log-likelihood function including the penalty matrix and the smoothing parameter μ

$$ll_{pen} = ll - \frac{1}{2}\,\mu \boldsymbol{\beta}^\top \mathbf{D_2}^\top \mathbf{D_2} \boldsymbol{\beta}, \tag{5.11}$$

where ll is the log-likelihood described in Section 5.2.1, that is extended by a penalty term (see e.g. Fahrmeir et al, 2007).

Illustrative example
The effect of the penalization on the basis function weights is illustrated in a small simulated example. Event times were generated for 3,000 individuals from an event time distribution with hazard function

$$\lambda(t) = 0.06 \left(1 + \frac{8}{\exp(0.2\,t)}\right).$$

Censoring times were assumed to follow an exponential distribution with a hazard rate of one and observation times were administratively censored at t=5.
The hazard function was estimated using the B-spline and the P-spline approach with five interior knots and basis functions as described in Equation 5.1. Equidistant knots as presented in Figure 5.1 (a) were used. The true hazard function and the unweighted basis functions are shown in Figure 5.2 (a). Weights for the B-spline basis functions were estimated by numerical maximization of the log-likelihood function shown in Equation 5.6 for the case without penalization and Equation 5.11 for consideration of a smoothing parameter. In Figure 5.2 (b) the weighted basis function and the estimated hazard function, which was derived as the sum of the weighted basis functions for each timepoint, are presented for μ=0. As no penalisation was considered, the resulting hazard function is rough due to large differences in nearby basis function weights. In picture (c) the weighted basis functions and the estimated hazard rate considering a smoothing parameter of μ=0.1 are shown. For Figure 5.2 (d) a smoothing parameter of μ=100 was chosen. As differences between nearby regression coefficients are penalized in Figures (c) and (d), weights are similar for neighbouring basis functions leading to less rough estimates of the hazard function, with a smoother hazard function obtained for the higher value of the smoothing parameter μ.

5.3.2 Extension to the mixture model approach

In this section the extension of the penalization described in Section 5.3.1 to a competing risks mixture model is described and discussed. For convenience the setting with a binary

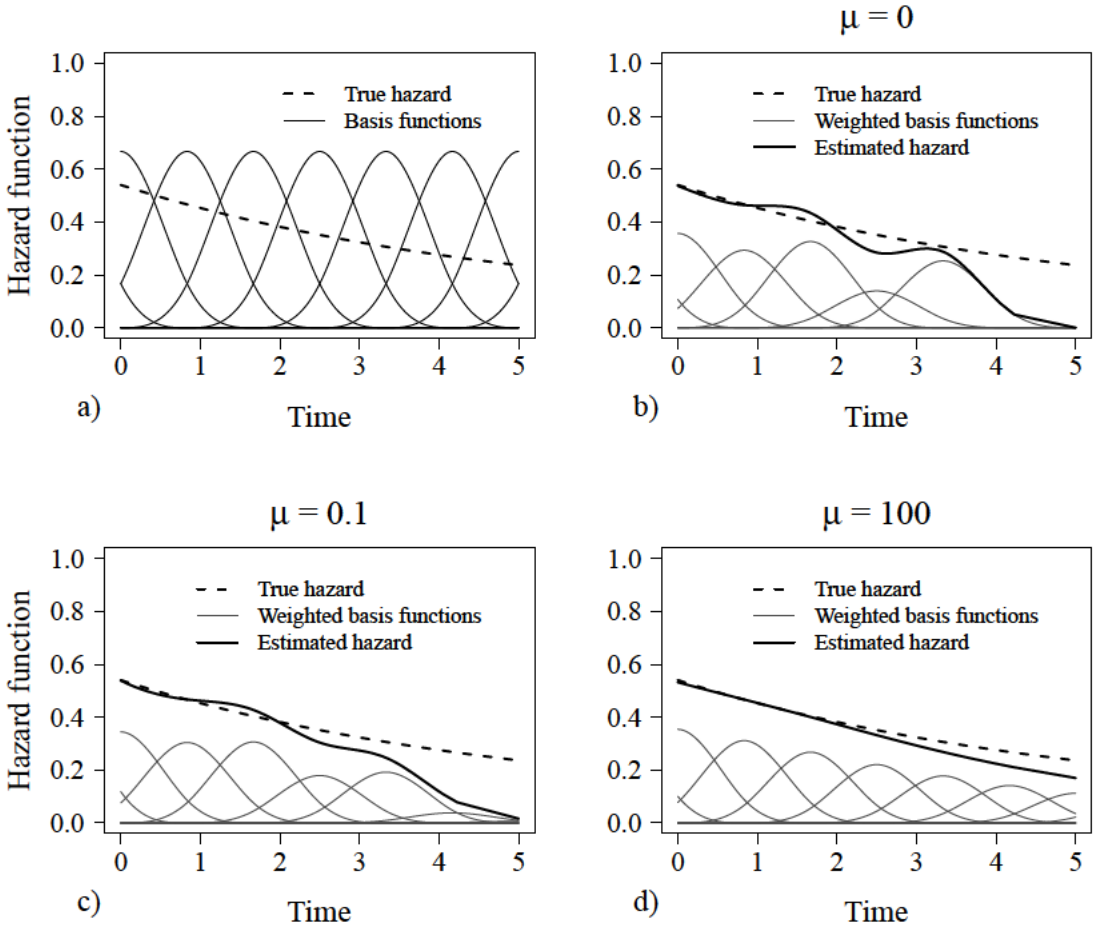

Figure 5.2: Predefined hazard rate and unweighted basis functions (a). Estimated hazard rate and weighted basis functions without penalisation (b). Estimated hazard rate and weighted basis functions for a smoothing parameter of μ=0.1 (c). Estimated hazard rate and weighted basis functions for a smoothing parameter of μ=100 (d).

covariate and two possible types of event is considered again. The conditional hazard rates of the mixture model are to be estimated using spline functions as shown in Equation 5.7. As the vector of regression coefficients $\boldsymbol{\beta}_{MM}$ consists of different components, namely the regression coefficients indicating covariate influence on the marginal event type probabilities and the vectors of regression coefficients weighting the basis functions for the different conditional hazard rates ($\boldsymbol{\beta}_{MM}$=$(\mu, \boldsymbol{\beta}_{1,0}, \boldsymbol{\beta}_{1,1}, \boldsymbol{\beta}_{2,0}, \boldsymbol{\beta}_{2,1})^\top$), the penalty matrix used in the penalized log-likelihood (Equation 5.11) has to be adapted. For the considered scenario the penalty matrix can be written as

$$\mathbf{D_{MM}} = \begin{pmatrix} \mathbf{0} & 0 & \dots & \dots & 0 \\ 0 & \mathbf{D_2}^\top\mathbf{D_2} & 0 & \dots & 0 \\ 0 & 0 & \mathbf{D_2}^\top\mathbf{D_2} & 0 & 0 \\ 0 & \dots & 0 & \mathbf{D_2}^\top\mathbf{D_2} & 0 \\ 0 & \dots & \dots & 0 & \mathbf{D_2}^\top\mathbf{D_2} \end{pmatrix}, \tag{5.12}$$

with $\mathbf{D_2}$ as defined in Equation 5.10. The matrix in the topleft corner will be a 2×2-matrix of zeros, as regression coefficients for the marginal event type distribution are not penalized. Each of the $\mathbf{D}_2$ matrices will be of dimension $(G+2) \times (G+4)$ and consequently the $\mathbf{D_2}^\top \mathbf{D_2}$ components will have dimension $(G+4) \times (G+4)$.
In order to estimate the regression coefficients, the log-likelihood including the penalty term will be maximized numerically. The log-likelihood of the mixture model approach using B-spline functions for estimation of conditional hazard rates, displayed in Equation 5.9 without penalization, can be extended as shown in Equation 5.11 replacing the penalty matrix $\mathbf{D_2}^\top \mathbf{D_2}$ by $\mathbf{D_{MM}}$, which is presented in Equation 5.12.

5.4 Discussion and outlook

In their article describing the estimation of cause-specific and subdistribution hazards from a mixture model, Lau et al (2011) recommend to assume, that the conditional hazard rates follow a flexible parametric survival distribution, as the three-parameter generalized gamma distribution, in order to evaluate time-dependence of the hazard rates and hazard ratios. The approach was applied to a real data set, but no simulation studies investigating the numerical stability of the approach and the ability to detect time-varying cause-specific and subdistribution hazard rates and hazard ratios were performed. As in a small simulation study (results not shown) the generalized gamma approach appeared to be numerically unstable, an alternative approach using penalized B-spline functions for estimation of conditional hazard rates in a mixture model is presented here. Both approaches are compared in a simulation study regarding their abilities to estimate time-constant and time-dependent cause-specific and subdistribution hazards and hazard ratios. The simulation study is presented in Section 7, the algorithms for generation of competing risks data are described in Section 6. The new approach was also applied to a real data set from a clinical cohort study investigating a risk stratification for cardiac death in patients who survived a myocardial infarction, with the aim to estimate cause-specific and subdistribution hazard ratios from a mixture model.
The proposed method using penalized spline functions for estimation of conditional hazard rates in a mixture model was also described in an article, which was submitted for publication and was under review when this work was finalized. The article was co-authored by Prof. Dr. Georg Schmidt from the 1. Medizinischen Klinik of the Klinikum rechts der Isar of the Technische Universität München and by Prof. Dr. Kurt Ulm of the Institut für Medizinische Statistik und Epidemiologie of the Technische Universität München. The article includes a description of the proposed method as shown in Sections 5.2.2 and 5.3.2, parts of the simulation study presented in Section 7 and results obtained from application of the method to the clinical data shown in Section 8.3.

Chapter 6

Simulation of competing risks data

As newly introduced statistical approaches are often of complex nature, including for example weighting schemes or resampling methods, an analytical evaluation of these methods and comparisons of different approaches are often not possible. So, in order to identify the best method for a given data situation, simulation studies are applied comparing relevant methods under different scenarios.
A simulation study relies on an adequate algorithm for generation of appropriate data. Several algorithms for data generation, including methods for simulation of data following various survival time distributions, are described by Gentle (2003). The generation of survival data for proportional hazards models was described by Leemis (1987) and Bender et al (2005). The simulation of competing risks data following prespecified cause-specific hazards was presented by Beyersmann et al (2009). It was also described how cause-specific hazards have to be selected to obtain data providing a given subdistribution hazard rate using the relationship between cause-specific and subdistribution hazards. In this section the algorithm for generation of competing risks data following cause-specific hazards and the choice of hazards to generate competing risks data following predefined subdistribution hazard rates as well as constraints for hazard rates considered for simulation are described. In Section 6.3 approaches for data generation with time-dependent cause-specific hazard rates, intended for generation of competing risks data with a given subdistribution hazard rate, using the inversion method or adapting a method for generation of time-to-event data with time-varying covariates originally presented by Sylvestre and Abrahamowicz (2008), are described. The methods introduced in this section were applied for data generation in the simulation study described in Section 7 for comparison of parametric mixture models and mixture models using the spline approach for estimation of conditional hazard rates introduced in Section 5.
Description of the Binomial algorithm for generation of competing risks data with a predefined subdistribution hazard rate for the event of interest, presented in Section 6.3.3, and constraints for choices of hazard rates (Section 6.3.2), as well as the simulation study validating the data generating procedure (Section 6.3.4) were published in the *Journal of Statistical Computation and Simulation* (Haller and Ulm, 2013)

6.1 Simulation of time-to-event-data using the inversion method

Leemis (1987) and Bender et al (2005) described, how time-to-event data depending on a vector of covariates $\mathbf{x}$ can be generated for proportional hazards models using the inversion method (see e.g. Gentle, 2003). For application of the inversion method a random number generator for uniformly distributed random numbers in the interval zero to one is needed. Event times are assumed to follow a parametric event time distribution fulfilling the proportional hazards assumption, as the exponential distribution or the Weibull distribution for a fixed shape parameter.
Generally, the inversion method can be applied for generation of random numbers following a predefined distribution. For event time data the distribution is mostly defined by its hazard or cumulative hazard function. Event times can be generated by solving the equation

$$U = F(T|\mathbf{x}) \tag{6.1}$$

or

$$U = S(T|\mathbf{x}), \tag{6.2}$$

with U being a random number following a uniform distribution in the interval $[0; 1]$. Defining the event time distribution by the hazard function, Equation 6.2 can also be denoted as

$$U = \exp\big(-\Lambda(T|\mathbf{x})\big) = \exp\left(-\int_0^T \lambda(s|\mathbf{x})\mathrm{d}s\right), \tag{6.3}$$

with $\lambda(t|\mathbf{x})$ being the hazard function and $\Lambda(t|\mathbf{x})$ being the cumulative hazard function.
For the exponential distribution, modelling the hazard rate via $\lambda(t|\mathbf{x}) = \exp(\boldsymbol{\beta}^\top\mathbf{x})$, event times can be generated following Equation 6.3 as

$$T = -\frac{\ln U}{\exp(\boldsymbol{\beta}^\top\mathbf{x})}. \tag{6.4}$$

Transformations of the uniformly distributed random number to generate event-time data following exponential distributions, Weibull distributions or Gompertz distributions are presented in the article by Bender et al (2005). Censored event times can be introduced by additional simulation of censoring times. For each individual the minimum of the generated event time and the censoring time will be considered as observed time and the status variable will be set to zero, if the censoring time was smaller than the event time, indicating a censored observation, or to one else.

6.2 Simulation of competing risks data following prespecified cause-specific hazards

Beyersmann et al (2009) presented an algorithm for the generation of competing risks data using predefined cause-specific hazards. The algorithm is also described in the textbook by Beyersmann et al (2012). As the cause-specific hazard rates are the forces that "completely

determine the competing risks process" (Beyersmann et al, 2009), a simulation of competing risks data using the cause-specific hazard rates appears to be the natural way. As for each individual the overall hazard at each timepoint is the sum of the cause-specific hazard rates for all K possible types of event, the event time is generated from an event time distribution with hazard rate $\lambda_{ov.}(t|\mathbf{x}) = \sum_{k=1}^{K} \lambda_k(t|\mathbf{x})$ in a first step. Then the type of event is determined by a Bernoulli experiment with the probabilities for each event type $k=\{1,\ldots,K\}$ being proportional to the cause-specific hazard rates $\lambda_k(t|\mathbf{x})$ at the drawn event time.
So competing risks data following predefined cause-specific hazards, that possibly depend on covariates, can be generated for n individuals as described in the following algorithm. For convenience the algorithm is presented for two possible types of event, i.e. K=2, but it can be easily adapted to more event types following the description above.

1. Define the cause-specific hazard rates $\lambda_1(t|\mathbf{x})$ and $\lambda_2(t|\mathbf{x})$ for both types of event e.g. by two Cox-type regression models with possibly time-dependent cause-specific (log-)hazard ratios $\lambda_k(t|\mathbf{x}) = \lambda_{k;0}(t)\exp\left(\boldsymbol{\beta_k}(t)^\top \mathbf{x}\right)$ for $k=\{1,2\}$.

2. Start with subject i=1.

3. Use an adequate procedure to generate an event time with overall hazard rate $\lambda_{ov.}(t|\mathbf{x_i}) = \lambda_1(t|\mathbf{x_i}) + \lambda_2(t|\mathbf{x_i})$ for individual i.

4. After simulation of the i^{th} event time t_i, determine the type of event by a Bernoulli experiment with probabilities p_1=$\lambda_1(t_i|\mathbf{x_i})/\left(\lambda_1(t_i|\mathbf{x_i})+\lambda_2(t_i|\mathbf{x_i})\right)$ for an event of type k=1 and p_2=$\lambda_2(t_i|\mathbf{x_i})/\left(\lambda_1(t_i|\mathbf{x_i})+\lambda_2(t_i|\mathbf{x_i})\right)$ for k=2.

5. Continue with 3. for individual i=1.

The event times in step 3 can be determined using an adequate random number generator if the event times follow a common survival distribution. The inversion method presented in Section 6.1 can be applied for any valid distribution defined by the overall hazard function, possibly using numerical procedures for calculation of the cumulative overall hazard function and for solution of Equation 6.3. An alternative approach for generation of event times following an arbitrary hazard function is presented in Section 6.3.3. For assignment of the event type in step 4, a uniformly distributed random number V can be drawn and an event of type 1 will be assigned, if $V<p_1$, and an event of type 2, else. Censoring times can be considered as described in Section 6.1.
Applications of that algorithm can e.g. be found in Beyersmann et al (2009), Grambauer et al (2010), Allignol et al (2011), and Beyersmann et al (2012).

6.3 Simulation of competing risks data following prespecified subdistribution hazards

In recent years different methods focussing on the subdistribution hazard, which was described in in Section 3.3.3 (Equation 3.11), were presented (e.g. Katsahian et al, 2006; Ruan and Gray, 2008; Sun et al, 2008). In order to evaluate the behaviour of these methods under different scenarios, as e.g. varying censoring schemes, or to investigate the robustness

of the methods under violation of model assumptions, simulation studies using adequately generated competing risks data have to be performed. For approaches focussing on the subdistribution hazards, competing risks data providing predefined subdistribution hazard rates have to be generated. In most articles discussing approaches, that use the subdistribution hazard, competing risks data were generated from a unit exponential mixture distribution (Fine and Gray, 1999), but this procedure, which is sketched in Section 6.3.1, does not allow to specify flexible subdistribution hazard rates directly. Beyersmann et al (2009) showed how cause-specific hazards, used for data generation following the algorithm presented in Section 6.2, have to be chosen to obtain competing risks data following the desired subdistribution hazard rates. In order to apply the approach, generation of event time data with time-dependent hazard rates has to be performed. The theoretical background presented by Beyersmann et al (2009) is summarized in Section 6.3.2 and certain constraints, that have to be fulfilled in order to obtain adequate competing risks data, are described. Generation of event times can be performed using the inversion method presented in Section 6.1. An alternative approach applying the Binomial Algorithm, originally proposed by Sylvestre and Abrahamowicz (2008) for simulation of time-to-event data with time-varying covariates, is presented in Section 6.3.3. In Section 6.3.4 the method using the Binomial Algorithm is validated for different scenarios using established methods for subdistribution hazards analysis in the competing risks setting.

6.3.1 Simulation using a unit exponential mixture distribution

In their article presenting the proportional subdistribution hazards regression model, Fine and Gray (1999) used a simulation approach for generation of competing risks data with a predefined subdistribution hazard ratio, which was later adapted by different authors (e.g. Latouche et al, 2004; Deslandes and Chevret, 2010). The cumulative incidence function for the event of interest (k=1) is defined as

$$\underline{F}_1(t|\mathbf{x}) = 1 - \Big(1 - p\,\big(1-\exp(-t)\big)\Big)^{\exp(\boldsymbol{\eta_1}^\top \mathbf{x})}, \tag{6.5}$$

with p being a predefined probability for an event of type k=1 for an individual with all covariates being equal to zero and $\boldsymbol{\eta_1}$ representing the vector of subdistribution log-hazard ratios for event type k=1.
For a covariate vector of $\mathbf{X}$=$\mathbf{0}$, representing the baseline situation, this is a unit exponential mixture model, as

$$\begin{aligned} \underline{F}_1(t|\mathbf{X}{=}\mathbf{0}) &= 1 - \Big(1 - p\big(1-\exp(-t)\big)\Big)^{\exp(\boldsymbol{\eta_1}^\top \mathbf{0})} = \\ &= 1 - \Big(1 - p\big(1-\exp(-t)\big)\Big) = \\ &= p\big(1-\exp(-t)\big) = p\,\overline{F}_1(t|\mathbf{X}{=}\mathbf{0}), \end{aligned} \tag{6.6}$$

with $\overline{F}_1(t|\mathbf{x})$ being the distribution function of an exponential model with a hazard rate of one.
Generation of competing risks data is performed following the mixture model definition by first drawing the type of event, using the individual's marginal event type probability,

and subsequently simulating the event time given the type of event. For individuals with a covariate vector unequal to zero, the event probabilities can be derived by calculating the value of the cumulative incidence function presented in Equation 6.5 for t going to infinity.

6.3.2 Using the relationship between cause-specific and subdistribution hazards

Cause-specific hazards needed for simulation

In the appendix of their article, Beyersmann et al (2009) presented how cause-specific hazards have to be chosen for data generation in a setting with two possible types of failure, in order to obtain competing risks data, that provide the desired subdistribution hazard for the event of interest. The approach is based on the relationship between the subdistribution hazard for the event of interest $\gamma_1(t|\mathbf{x})$ and the cause-specific hazard rates $\lambda_1(t|\mathbf{x})$ and $\lambda_2(t|\mathbf{x})$ presented in Beyersmann and Schumacher (2007) and shown in Equation 3.14 of this work. For data generation, two of the three relevant measures, possibly depending on a covariate vector $\mathbf{x}$, namely the subdistribution hazard for the event of interest $\gamma_1(t|\mathbf{x})$, the cause-specific hazard for the event of interest $\lambda_1(t|\mathbf{x})$ and the cause-specific hazard for the competing event $\lambda_2(t|\mathbf{x})$, have to be chosen. The third measure is derived from the given two using the relationship between cause-specific and subdistribution hazards.
When the desired subdistribution hazard and the cause-specific hazard for the event of interest are defined, the cause-specific hazard for the competing event can be derived as

$$\lambda_2(t|\mathbf{x}) = \gamma_1(t|\mathbf{x}) - \lambda_1(t|\mathbf{x}) - \frac{\mathrm{d}}{\mathrm{d}t}\ \ln\left(\frac{\gamma_1(t|\mathbf{x})}{\lambda_1(t|\mathbf{x})}\right). \tag{6.7}$$

When $\lambda_2(t|\mathbf{x})$ and $\gamma_1(t|\mathbf{x})$ are defined, the cause-specific hazard for the event of interest can be determined using

$$\lambda_1(t|\mathbf{x}) = \frac{\gamma_1(t|\mathbf{x})\ \exp\big(-\Gamma_1(t|\mathbf{x}) + \Lambda_2(t|\mathbf{x})\big)}{1 - \int_0^t \gamma_1(s|\mathbf{x})\ \exp\big(-\Gamma_1(s|\mathbf{x}) + \Lambda_2(s|\mathbf{x})\big)\mathrm{d}s}. \tag{6.8}$$

It can be seen easily from Equation 3.14 that determination of both cause-specific hazards leads to a subdistribution hazard of the form

$$\gamma_1(t|\mathbf{x}) = \lambda_1(t|\mathbf{x}) \Big/ \left(1 + \frac{F_2(t|\mathbf{x})}{S_{ov.}(t|\mathbf{x})}\right). \tag{6.9}$$

The derivations of these equations are described in more detail in Beyersmann et al (2009) and Beyersmann et al (2012).
The cause-specific hazards $\lambda_1(t|\mathbf{x})$ and $\lambda_2(t|\mathbf{x})$, either predefined or derived from one of the Equations 6.7 or 6.8, can be used for generation of competing risks data following the algorithm described in Section 6.2. As the obtained hazard functions will generally result in a time-dependent overall hazard function, which is necessary for simulation of event times, adequate methods for data generation are inevitable. Event times can either be generated using the inversion method presented in Section 6.1, which can be applied using arbitrary cumulative hazard functions, or using an algorithm for simulation of event time data with time-dependent hazard rates, which is based on the Binomial Algorithm by Sylvestre and Abrahamowicz (2008) and is presented in Section 6.3.3.

Constraints for hazard functions

In order to obtain event times following the desired prespecified subdistribution hazards, the following constraints have to be fulfilled:

- All hazard functions have to be non-negative for all time points $t > 0$.
- In the presence of competing risks, i.e. when $\lambda_2(t|\mathbf{x})>0$ for any timepoint t, the cumulative incidence function for event type $k=1$, that is related to the cumulative subdistribution hazard through $\underline{F}_1(t|\mathbf{x}) = 1 - \exp\big(-\Gamma_1(t|\mathbf{x})\big)$, has to converge to $P(D=1)$ for t going to infinity. $\gamma_1(t|\mathbf{x})$ has to be chosen accordingly, and consequently $\gamma_1(t|\mathbf{x})$ has to converge to zero and $\Gamma_1(t|\mathbf{x})$ must not converge to infinity for t approaching infinity.
- For t going to zero, $\lambda_1(t|\mathbf{x})$ and $\gamma_1(t|\mathbf{x})$ have to converge to the same value, as cause-specific and subdistribution hazards are identical before occurrence of the first competing event.

6.3.3 The Binomial Algorithm for simulation of competing risks data with time-dependent hazard rates

Time-to-event data following time-dependent hazard rates, necessary for simulation of competing risks data providing predefined subdistribution hazards, can be generated using the inversion method presented in Section 6.1, but this can be time-consuming, as numerical approaches might be necessary for calculation of the cumulative hazard rate and for determination of the event time following Equation 6.3 for each individual. An alternative approach, which is based on the Binomial Algorithm of Sylvestre and Abrahamowicz (2008), that was originally introduced for simulation of time-to-event data in the presence of time-varying covariates, is proposed in this Section. Event time data are generated for discrete timepoints considering the overall hazard rate as conditional event probability. The approach is also adapted to the competing risks setting for generation of event time data with mutually exclusive types of event following a given subdistribution hazard rate for the event of interest. The algorithm for data generation was evaluated for different settings in a simulation study, which is presented in Section 6.3.4.

The Binomial Algorithm for generation of event time data following time-dependent hazard rates

The adapted algorithm of Sylvestre and Abrahamowicz for generation of time-to-event data with time-dependent hazard rates and one possible endpoint can be denoted as follows:

1. Define the hazard rate possibly depending on a vector of covariates $\lambda(t|\mathbf{x})$.
2. Start with individual $i=1$.
3. Begin at time $t=1$.

4. Determine the conditional probability of subject i for failing at time t, given it survived up to time t, depending on its covariate information $\mathbf{x_i}$, following the model defined in the first step: $p(t|\mathbf{x_i}) = \lambda(t|\mathbf{x_i})$.

5. Draw a random number to determine, whether individual i fails at time t, e.g. by drawing a uniform random number and comparing it to $p(t|\mathbf{x_i})$ or by using an adequate sampling mechanism.

6. If i was determined not to fail at time t, go on with 4. for time t+1, else go on with 3. for individual i+1.

Adapting the Binomial Algorithm for simulation of competing risks data following given subdistribution hazards rates

The method described before can be used to generate event time data with time-dependent overall hazard, which is necessary for simulation of competing risks data following predefined subdistribution hazards using the method proposed by Beyersmann et al (2009) and described in Section 6.3.2. The complete algorithm for generation of competing risks data following predefined subdistribution hazard rates can be denoted as follows:

1. Define the subdistribution hazard rate $\gamma_1(t|\mathbf{x})$ and one cause-specific hazard rate $\lambda_k(t|\mathbf{x})$, depending on the vector of covariates $\mathbf{x}$, and determine the other cause-specific hazard rate following Equation 6.7 or 6.8, so that constraints described in Section 6.3.2 are fulfilled for all hazard rates.

2. Start with subject i=1.

3. Begin at time t=1.

4. Define the conditional probability of subject i to fail from any cause at time t, given survival up to time t, as $p(t|\mathbf{x_i}) = \lambda_1(t|\mathbf{x_i})+\lambda_2(t|\mathbf{x_i})$.

5. Draw a random number to determine if the individual failed at time t, e.g. by drawing a uniformly distributed random number V and assigning an event if $V < p(t|\mathbf{x_i})$.

6. If an event was observed, the type of event is determined by a Bernoulli experiment with probabilities $\lambda_1(t|\mathbf{x_i})/\big(\lambda_1(t|\mathbf{x_i})+\lambda_2(t|\mathbf{x_i})\big)$ for an event of type 1 and $\lambda_2(t|\mathbf{x_i})/\big(\lambda_1(t|\mathbf{x_i})+\lambda_2(t|\mathbf{x_i})\big)$ for an event of type 2.

7. If t was determined to be the event time for individual i go on with 3. for individual i+1, else return to 4. for time t+1.

Censoring times can be considered as described in Section 6.1. The behaviour of the algorithm for generation of competing risks data was validated for different scenarios. The data generating algorithm and the results of the validation were published in Haller and Ulm (2013).

6.3.4 Validating the data generating process

Different scenarios were chosen to validate the algorithm for generation of competing risks data following prespecified subdistribution hazards using the Binomial Algorithm as described in Section 6.3.3:

- One population - no covariates
- Two groups - time-constant subdistribution hazard ratio
- Two groups - time-varying subdistribution hazard ratio
- One quantitative covariate
- A multiple regression model

Established standard methods for subdistribution hazards analysis of competing risks data were used. As the focus of the simulation study is on behaviour of the data generating process, not on the methods used for analysis, no censored observations were considered.

Example 1: One population - no covariates

Firstly, event times following a predefined subdistribution hazard rate were generated for one population without consideration of covariate effects, i.e. all subjects follow the same predefined hazard function, which is independent of the covariate vector $\mathbf{x}$. The subdistribution hazard $\gamma_1(t)$ and the cause-specific hazard for the event of interest $\lambda_1(t)$ were specified. The subdistribution hazard was chosen to lead to an expected proportion of events of interest of one half, if each subject will be observed until it failed from one of the two possible events. The chosen subdistribution and cause-specific hazard rates for the event of interest and the corresponding cumulative hazards are

$$\begin{aligned}
\gamma_1(t) &= 0.001 \ \exp\left(-\frac{0.001\,t}{\ln(2)}\right) \\
\lambda_1(t) &= 0.001 \\
\Gamma_1(t) &= \ln(2)\left(1-\exp\left(-\frac{0.001\,t}{\ln(2)}\right)\right) \\
\Lambda_1(t) &= 0.001\,t
\end{aligned}$$

According to Equation 6.8, the cause-specific hazard for the competing event, which is necessary for data generation following the proposed algorithm, was determined to be

$$\lambda_2(t) = 0.001 \ \exp\left(-\frac{0.001\,t}{\ln(2)}\right) - 0.001 + \frac{0.001}{\ln(2)}.$$

Since $\lambda_1(t)$, $\lambda_2(t)$ and $\gamma_1(t)$ are strictly positive functions for all $t>0$, $\lambda_1(t)$ converges to $\gamma_1(t)$ for t approaching zero and $\Gamma_1(t)$ converges to $\ln(2)$ for t going to infinity, translating to a cumulative incidence function converging to one half, all restrictions presented in Section 6.3.2 are met. In Figure 6.1 (a) the cause-specific hazard rates, used for generation of

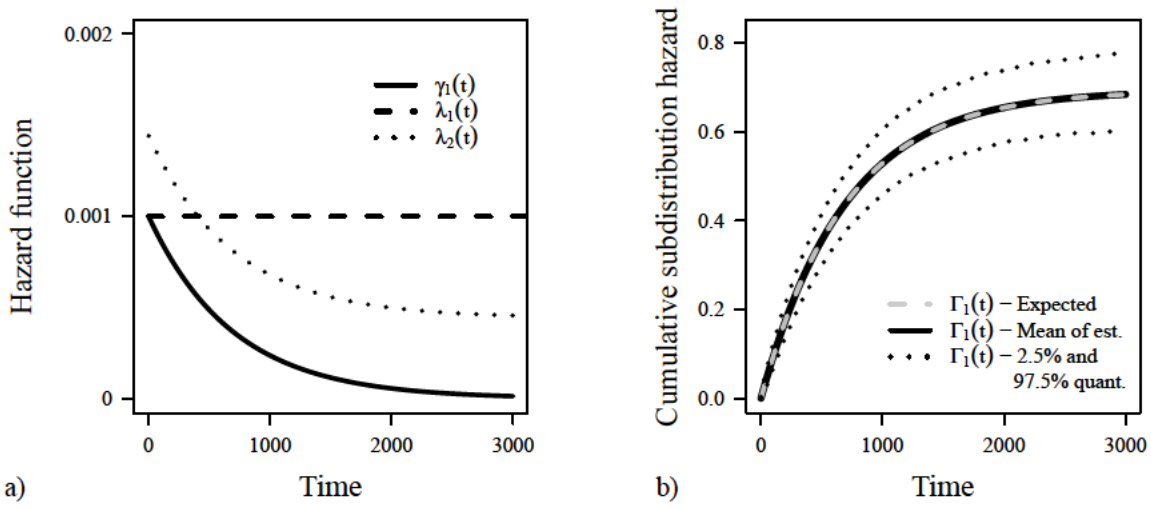

Figure 6.1: Example 1 – Left: Cause-specific hazards (dashed and dotted lines) used for simulation to obtain the desired subdistribution hazard (solid). Right: Mean of the estimated cumulative subdistribution hazards (solid line) from generated data and corresponding 2.5% and 97.5% quantiles assessed for each timepoint. The dashed grey line represents the expected value.

competing risks data following the desired subdistribution hazard, and the subdistribution hazard itself are shown. It can be seen that the cause-specific hazard for the competing event is time-dependent. Competing risks data were simulated following the algorithm described in Section 6.3.3 using the cause-specific hazards $\lambda_1(t)$ and $\lambda_2(t)$. Four thousand datasets with 500 observations each were generated.
For every dataset an estimate $\hat{\gamma}_1(t)$ for the subdistribution hazard $\gamma_1(t)$ was derived for each possible event time. As no censoring was present, the subdistribution hazard for a given time t was estimated as the number of events of interest at t divided by the number of subjects in the modified risk set defined in Equation 3.11 (see e.g. Putter et al, 2007, and Equation 3.16). In Figure 6.1 (b) the mean of the estimated cumulative subdistribution hazard functions from 4,000 generated datasets is shown for each timepoint considered (solid line). The expected cumulative subdistribution hazard function $\Gamma_1(t)$ is displayed as dashed grey line. The 2.5% and 97.5% quantiles are presented to indicate variability

Time	$t = 50$	$t = 100$	$t = 200$	$t = 500$	$t = 1000$	$t = 2000$
$\mathbf{\Gamma_1(t)}$	**0.048**	**0.093**	**0.174**	**0.356**	**0.529**	**0.654**
Mean	0.048	0.093	0.174	0.356	0.529	0.655
Median	0.047	0.092	0.174	0.356	0.527	0.653
Std.dev.	0.010	0.014	0.019	0.029	0.037	0.042
$Q_{.025}$	0.030	0.066	0.137	0.301	0.458	0.575
$Q_{.975}$	0.068	0.123	0.213	0.415	0.604	0.741

Table 6.1: Example 1 – Summary of the estimated cumulative subdistribution hazard functions at different timepoints. Expected values are printed in bold font. 4,000 datasets with 500 observations were generated.

of the estimated cumulative subdistribution hazard rates between the generated datasets. Summaries of estimates for the cumulative subdistribution hazard function obtained in the 4,000 simulation runs are shown for different timepoints in Table 6.1. The expected values of the cumulative subdistribution hazard function are printed in bold font. Both, Figure 6.1 and Table 6.1, indicate very good agreement between the expected cumulative subdistribution hazard function and the mean cumulative subdistribution hazard function estimated from generated data.

Example 2: Two group comparison - constant subdistribution hazard ratio

When competing risks data for two independent groups are desired, data can be generated separately for both groups. A binary covariate X is introduced to indicate group membership. The group with X=0 will be called reference group throughout the section, X=1 indicates the so called study group. For each group the desired subdistribution hazard has to be specified and the cause-specific hazards for both types of event have to be chosen adequately, so that application of the data generating algorithm will lead to competing risks data providing the desired subdistribution hazards. The situation can also be described by a Cox-type regression model. The baseline hazards $\lambda_{k;0}(t)$ and $\gamma_{1;0}(t)$ denote the according hazards for the study group and the possibly time-dependent hazard ratios $\exp(\beta_k(t))$ and $\exp(\eta_1(t))$, respectively, are defined as the ratio of the cause-specific hazards for event k or the subdistribution hazards for the event of interest between the study group (X=1) and the reference group (X=0). For each group, two of the three measures under consideration, the cause specific hazards for both event types $\lambda_1(t|x)$ and $\lambda_2(t|x)$ and the subdistribution hazard for the event of interest $\gamma_1(t|x)$, have to be specified, while the third measure is calculated from these two following one of Equations 6.7 to 6.9. All hazard rates for both groups - predetermined or calculated - have to fulfil the constraints presented in Section 6.3.2, so the hazard rates considered have to be chosen with caution.
In this example, the data should provide a predetermined subdistribution hazard ratio $\exp(\eta_1)$ between both groups, which is constant over time. The subdistribution hazards for the event of interest $\gamma_1(t|x)$ and the cause-specific hazards for the competing event $\lambda_2(t|x)$ were specified, so $\lambda_1(t|x)$ had to be calculated from Equation 6.8 for both groups. As the subdistribution hazard functions for both groups are constrained to a form that leads to cumulative incidence functions converging to a value smaller than one for t going to infinity, the choice of these functions is not straightforward. Here, the baseline subdistribution hazard, i.e. the subdistribution hazard for the group with X=0, was chosen to be

$$\gamma_{1;0}(t) = \gamma_1(t|X{=}0) = 0.001\ \exp\left(-\frac{0.001\,t}{\ln(1.5)}\right).$$

This leads to a cumulative incidence function for k=1 converging to one third for t going to infinity. Hence, in the reference group one out of three subjects is expected to fail from the cause of interest, two out of three from the competing event. The subdistribution hazard ratio was set to be 2, translating into a regression coefficient of $\eta_1 = \ln(2)$ and a subdistribution hazard for the study group of

$$\gamma_1(t|X{=}1) = 0.001\ \exp\left(-\frac{0.001\,t}{\ln(1.5)}\right)\ \exp\big(\ln(2)\big) = 0.002\ \exp\left(-\frac{0.001\,t}{\ln(1.5)}\right),$$

and so to a cumulative incidence function $\underline{F}_1(t|X{=}1)$ converging to a value of $5/9 \approx 0.556$ for t going to infinity. The cause-specific hazard functions for the competing event were chosen to be constant over time and equal for both groups, i.e. the risk of failing from event 2 is identical for individuals from both groups at each timepoint:

$$\lambda_2(t|X{=}0) = 0.001$$
$$\lambda_2(t|X{=}1) = 0.001$$

The cause-specific hazard functions for the event of interest $\lambda_1(t|X{=}0)$ and $\lambda_1(t|X{=}1)$ were calculated numerically for each point in time following Equation 6.8. The cause-specific hazards and the subdistribution hazards for the event of interest $k{=}1$ are illustrated for both groups in Figure 6.2 (a). The predefined subdistribution hazards and the calculated cause-specific hazards for the event of interest both decline over time. While the subdistribution hazard ratio was set to be constant for all timepoints, the ratio of cause specific hazard rates for the event of interest $\lambda_1(t|X{=}1)/\lambda_1(t|X{=}0)$ increases over time.

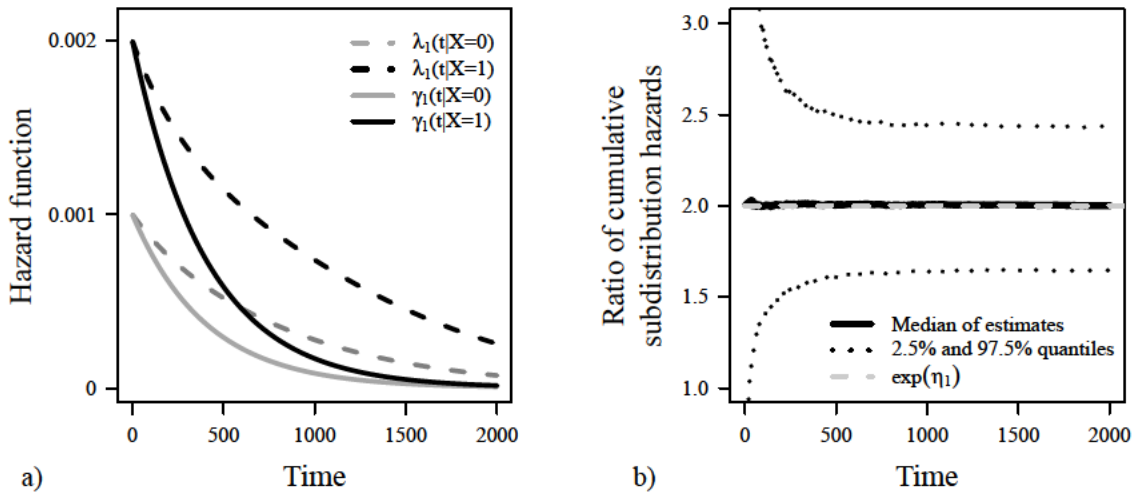

Figure 6.2: Example 2 – Left: Cause-specific (dashed lines) and subdistribution hazards (solid lines) for the event of interest for the reference (grey) and the study group (black). Right: Median ratio of the cumulative subdistribution hazards for each considered timepoint (solid black line) from 4,000 simulation runs with 1,000 observations each and corresponding 2.5% and 97.5% quantiles (dotted lines). The predefined hazard ratio is presented as dashed grey line.

Data were simulated following the algorithm presented in Section 6.3.3. 4,000 datasets were generated, each with 1,000 observations (500 per group). For each dataset the ratio of cumulative subdistribution hazards was estimated for every possible event time. In Figure 6.2 (b), the median of these ratios for each point in time as well as 2.5% and 97.5% quantiles of the estimates are presented. The dashed grey line indicates the expected subdistribution hazard ratio of $\exp(\eta_1){=}2$. As the ratio of the cumulative subdistribution hazards is shown, variability of the estimates decreases over time, as more information is considered for later timepoints.

In Table 6.2 results for simulation scenarios performed analogously to the one described above, but with different prespecified subdistribution hazard ratios $\exp(\eta_1)$, are presented. Logarithms of the hazard ratios were estimated fitting a proportional subdistribution hazards regression model as proposed by Fine and Gray (1999). Estimation was conducted using the R function *crr* from the library *cmprsk* (Gray, 2010). Means, medians, standard deviations, and 2.5% and 97.5% quantiles of the estimated regression coefficients are shown. The standard error provided by the function *crr* was used to estimate a 95% confidence interval for each simulation run, assuming normality for $\hat{\eta}_1$. The proportion of estimated confidence intervals that include the true value, denoted as *CI coverage*, are presented in the last row of Table 6.2. The summaries obtained from generated data reveal good agreement with the preset subdistribution log-hazard ratios, shown in bold font in the first row, for all values investigated.

η_1	**ln (0.8) = -0.223**	**ln (0.9) = -0.105**	**ln (1) = 0.000**	**ln (1.25) = 0.223**	**ln (1.5) = 0.405**	**ln (2) = 0.693**
Mean	-0.221	-0.108	-0.000	0.223	0.403	0.696
Median	-0.223	-0.107	-0.002	0.223	0.404	0.695
Std.dev.	0.115	0.110	0.110	0.105	0.103	0.099
$Q_{.025}$	-0.444	-0.324	-0.219	0.017	0.204	0.508
$Q_{.975}$	0.009	0.109	0.211	0.426	0.603	0.892
CI coverage	0.950	0.953	0.951	0.951	0.950	0.949

Table 6.2: Example 2 – Summaries of simulation scenarios with different prespecified subdistribution log-hazard ratios. 4,000 datasets, each with 500 observations per group, were generated per scenario. The Fine and Gray regression approach was conducted to estimate subdistribution log-hazard ratios.

Example 3: Two group comparison - time-dependent subdistribution hazard ratio

With an appropriate choice of cause-specific hazard functions, competing risks data following a time-dependent subdistribution hazard ratio can be generated. Using a Cox-type regression model for the subdistribution hazard, choice of the subdistribution hazard for the reference group $\gamma_1(t|X{=}0)$ and the time-dependent subdistribution hazard ratio $\exp(\eta_1(t))$ define the subdistribution hazard of the study group $\gamma_1(t|X{=}1)$.
For convenience the subdistribution hazard function for the reference group was chosen as in Example 2, but the hazard ratio was defined to be time-dependent

$$\exp\big(\eta_1(t)\big) = 1 + \frac{1}{\exp(0.001\,t)},$$

leading to a subdistribution hazard for the study group of

$$\begin{aligned}\gamma_1(t\,|X{=}1) &= \gamma_1(t\,|X{=}0)\ \left(1 + \frac{1}{\exp(0.001\,t)}\right)\\ &= 0.001\ \exp\left(-\frac{0.001\,t}{\ln(1.5)}\right)\ \left(1 + \frac{1}{\exp(0.001\,t)}\right).\end{aligned}$$

The hazard ratio converges to two for t approaching zero, but decreases to one for t going to infinity, indicating a difference in the subdistribution hazards between both groups for early timepoints that diminishes over the course of time. The cumulative subdistribution hazards for both groups are therefore

$$\Gamma_1(t\,|X{=}0) = -\ln(1.5)\,\exp\left(-\frac{0.001\,t}{\ln(1.5)}\right) + \ln(1.5)$$

and

$$\Gamma_1(t|X{=}1) = -\frac{\ln(1.5)}{1+\ln(1.5)}\left[\exp\left(-\frac{0.001\,t}{\ln(1.5)}\right) + \right.$$

$$\left. + \ln(1.5)\,\exp\left(-\frac{0.001\,t}{\ln(1.5)}\right) + \exp\left(-0.001\,t\,\frac{1+\ln(1.5)}{\ln(1.5)}\right) - 2 - \ln(1.5)\right],$$

leading to cumulative incidence functions for the event of interest converging to one third for $X{=}0$ and about one half (50.04%) for $X{=}1$ for t going to infinity.
The cause-specific hazard rates for the event of interest were specified for both groups. To fulfil the constraints, the cause-specific hazards for $k{=}1$ have to approach the same values as $\gamma_1(t|X{=}0)$ and $\gamma_1(t|X{=}1)$ for t going to zero, which are 0.001 and 0.002, respectively. For convenience, time-constant cause-specific hazard rates were chosen, so that $\lambda_1(t|X{=}0) = 0.001$ and $\lambda_1(t|X{=}1) = 0.002$. The cause-specific hazard rates for the competing event $\lambda_2(t|X{=}0)$ and $\lambda_2(t|X{=}1)$ were calculated following Equation 6.7. The cause-specific hazard functions and the subdistribution hazard function for the reference group are shown in Figure 6.3 (a), the corresponding functions for the study group are illustrated in Figure 6.3 (b).

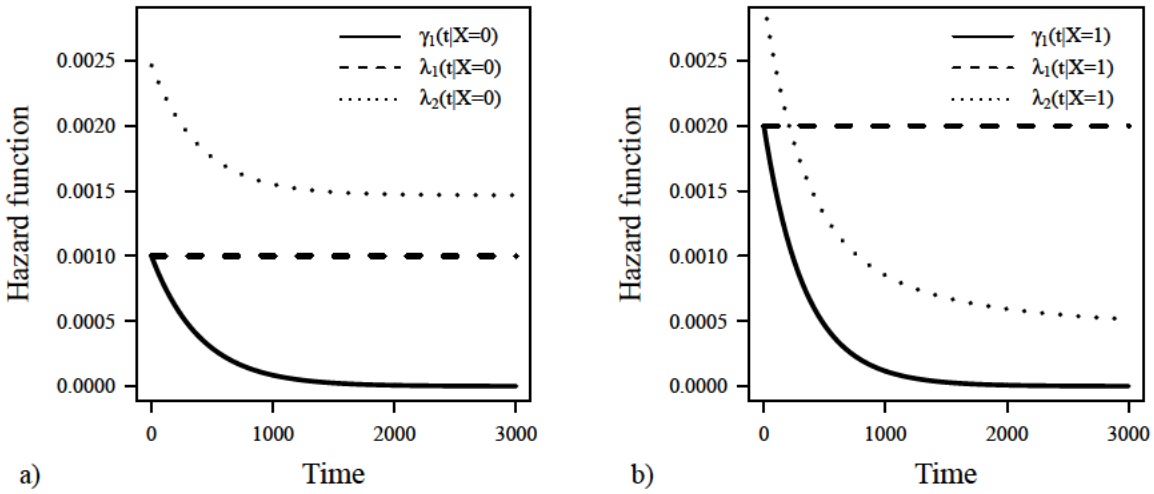

Figure 6.3: Example 3 – Desired subdistribution hazards (solid lines) for both groups (left: $X{=}0$, right: $X{=}1$), leading to a time-dependent subdistribution hazard ratio. Cause-specific hazard functions for both types of event, necessary to obtain the predefined subdistribution hazards, are shown for both groups by dashed ($k{=}1$) or dotted ($k{=}2$) lines.

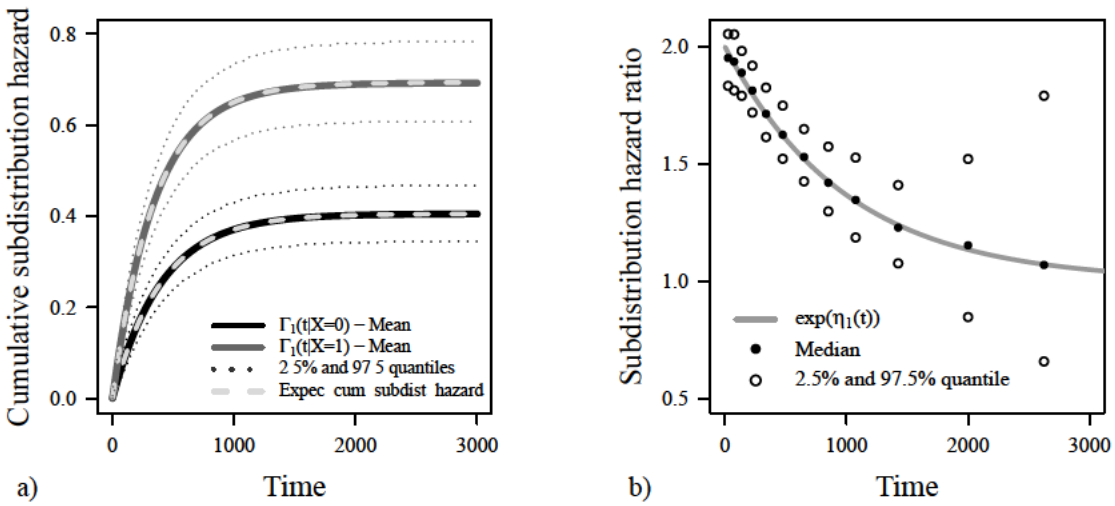

Figure 6.4: Example 3 – Left: Expected cumulative subdistribution hazards (dashed lines) following the model with time-dependent subdistribution hazard ratio and means (solid lines) and 2.5% and 97.5% quantiles (dotted lines) of the estimated cumulative subdistribution hazards for both groups (4,000 datasets with 500 observations per group, each). Right: The desired time-dependent subdistribution hazard ratio (grey line) and medians (filled dots) and 2.5% and 97.5% quantiles (open dots) of subdistribution hazard ratio estimates in prespecified time intervals from 4,000 simulation runs with 10,000 observations, each.

The simulation algorithm was conducted using these cause-specific hazard rates to generate 4,000 datasets with 500 observations per group, each. Means of the cumulative subdistribution hazard estimates for both groups at each considered timepoint, the 2.5% and 97.5% quantiles, and the expected values are displayed in Figure 6.4 (a). Summary statistics for estimates of the cumulative subdistribution hazard rates for both groups at certain timepoints are shown in Table 6.3.
To illustrate time-dependence of $\exp(\eta_1(t))$, the subdistribution hazard ratio between both groups was estimated for discrete time intervals. In order to investigate a substantial

Time	$t = 200$		$t = 500$		$t = 1000$		$t = 2000$	
Group	X=0	X=1	X=0	X=1	X=0	X=1	X=0	X=1
$\mathbf{\Gamma_1(t\|x)}$	**0.158**	**0.302**	**0.287**	**0.525**	**0.371**	**0.651**	**0.403**	**0.691**
Mean	0.157	0.302	0.286	0.525	0.370	0.650	0.401	0.690
Median	0.158	0.301	0.285	0.523	0.368	0.649	0.400	0.692
Std.dev.	0.018	0.026	0.026	0.038	0.030	0.044	0.032	0.047
$Q_{.025}$	0.121	0.253	0.238	0.446	0.317	0.565	0.342	0.600
$Q_{.975}$	0.193	0.353	0.334	0.600	0.430	0.737	0.465	0.784

Table 6.3: Example 3 – Summary of cumulative subdistribution hazard estimates for both groups at different timepoints for 4,000 simulation runs with 500 observations per group, each. A time-dependent subdistribution hazard ratio was considered. Expected values are shown in bold font.

number of time intervals and to observe an adequate number of events of interest in each of these intervals, new datasets with 10,000 observations, each, were generated using the same cause-specific hazard rates as above. The results obtained from 4,000 simulation runs are displayed in Figure 6.4 (b). The medians of estimated hazard ratios in each interval are shown as filled dots at the center of the interval. The corresponding 2.5% and 97.5% quantiles are displayed as open dots. The solid line represents the expected subdistribution hazard ratio over the course of time decreasing from two to one.
Again, the results presented in the figures and in the table indicate good agreement between the estimates obtained from the generated data and the expected values of the cumulative subdistribution hazard functions for both groups or the subdistribution hazard ratio, respectively.

Example 4: Quantitative covariate

Analogously to the proceeding described before for a binary covariate, competing risks data can be generated considering a quantitative covariate. As common for quantitative covariates the subdistribution hazard ratio does now, given the effect of the covariate on the subdistribution log-hazard rate is linear, denote the subdistribution hazard ratio between two individuals that differ in one unit of the investigated covariate. Special care has to be taken to fulfil the constraints, as validity of the model might also depend on the range of the covariate. Specification of the baseline situation, the subdistribution hazard ratio and the cause-specific hazard ratio for one type of event can result in a hazard ratio for the other event type that depends on time t and the covariate x. This is illustrated in the following example.
A Cox-type proportional subdistribution hazards regression model, implying a regression coefficient η_1, that is constant over time, is considered. The covariate x was chosen to follow a standard normal distribution. The baseline subdistribution hazard was preset to be

$$\gamma_{1;0}(t) = \gamma_1(t|X{=}0) = 0.001 \ \exp\left(-\frac{0.001\ t}{\ln(2)}\right),$$

leading to a cumulative incidence function converging to one half for t going to infinity for an individual with a covariate value of zero. The subdistribution hazard depending on covariate x is defined by

$$\gamma_1(t|x) = \gamma_{1;0}(t)\ \exp(\eta_1 x) = 0.001\ \exp\left(-\frac{0.001\ t}{\ln(2)}\right)\ \exp(\eta_1 x).$$

Different subdistribution hazard ratios $\exp(\eta_1)$, representing the quotient of the subdistribution hazards for two individuals differing in the covariate x by one unit, of 0.8, 0.9, 1, 1.25, and 1.5 were investigated. To ensure proper hazard functions for all possible covariate values, the range of x was restricted to the interval [-3; 3].
The cause-specific baseline hazard for the event of interest was chosen to be constant over time. To fulfil the restrictions presented in Section 6.3.2, it was set to be

$$\lambda_{1;0}(t) = \lambda_1(t|X{=}0) = 0.001.$$

The cause-specific hazard for the event of interest depending on covariate x was chosen to approach the cause-specific baseline hazard for t going to infinity and - as required - the

η_1	**ln (0.8) = -0.223**	**ln (0.9) = -0.105**	**ln (1) = 0.000**	**ln (1.25) = 0.223**	**ln (1.5) = 0.405**
Mean	-0.224	-0.106	-0.001	0.222	0.406
Median	-0.224	-0.106	-0.000	0.222	0.405
Std.dev.	0.045	0.045	0.045	0.045	0.046
$Q_{.025}$	-0.313	-0.194	-0.093	0.135	0.319
$Q_{.975}$	-0.137	-0.016	0.088	0.310	0.496
CI coverage	0.952	0.945	0.942	0.952	0.952

Table 6.4: Example 4 – Summaries of estimated subdistribution log-hazard ratios for different scenarios with one quantitative covariate. For each scenario 4,000 datasets with 1,000 observations were generated. The prespecified subdistribution log-hazard ratios are printed in bold font.

subdistribution hazard $\gamma_1(t|x)$ for t going to zero

$$\lambda_1(t|x) = \lambda_{1;0}(t)\, \exp\big(\eta_1 x\, \exp(-0.001\, t)\big).$$

The cause-specific hazard for the competing event $\lambda_2(t|x)$, necessary for the simulation of competing risks data following the desired subdistribution hazard, depends on time and the covariate value, and was therefore derived numerically for each timepoint following Equation 6.7, using the individual's covariate value.
For each of the five predefined subdistribution hazard ratios 4,000 datasets with 1,000 observations were generated. In each dataset the regression coefficient η_1 was estimated using a proportional subdistribution hazards model as proposed by Fine and Gray (1999). Analyses of the simulated data sets were conducted using the R function *crr* again. Summaries of estimated regression coefficients are shown in Table 6.4. Confidence interval coverage was assessed as described for Example 2. The obtained results indicate that the simulated data follow the desired models.

Example 5: Multiple subdistribution hazards regression

Finally, data for a multiple subdistribution hazards regression model were generated. A Cox-type regression model with a subdistribution baseline hazard $\gamma_{1;0}(t)$, describing the subdistribution hazard for a (possibly fictitious) individual with all covariate values equal to zero, and subdistribution hazard ratios for all covariates considered was defined. In order to generate the desired competing risks data, cause-specific hazard rates have to be chosen adequately so that Equation 3.14 holds for all $\mathbf{x}$ and t and well-behaved hazard functions for each possible combination of covariates are obtained.
As common in multiple regression models, the p^{th} regression coefficient can be interpreted as logarithm of the subdistribution hazard ratio for the event of interest between two subjects differing by one unit in the p^{th} covariate x_p and being identical in all other covariates $x_q, q \neq p$.
The baseline subdistribution hazard was set to be

$$\gamma_{1;0}(t) = \gamma_1(t|\mathbf{X}{=}\mathbf{0}) = 0.001\, \exp\left(-\frac{0.001\, t}{\ln(1.3)}\right),$$

leading to a proportion of expected type one and type two events of about 23% and 77% for an individual with all covariates equalling zero. The cause-specific baseline hazard for the event of interest was again set to be constant over time

$$\lambda_{1;0}(t) = \lambda_1(t|\mathbf{X}{=}\mathbf{0}) = 0.001,$$

leading to a cause-specific baseline hazard rate for the competing event, following Equation 6.7, of

$$\lambda_{2;0}(t) = \lambda_2(t|\mathbf{X}{=}\mathbf{0}) = 0.001 \exp\left(-\frac{0.001\,t}{\ln(1.3)}\right) + \frac{0.001}{\ln(1.3)} - 0.001.$$

Three covariates were considered in the simulation scenario, a continuous covariate X_1, which is uniformly distributed on the interval 0 to 3, a normally distributed covariate X_2 with mean 5 and variance 1, and a binary variate X_3 with a probability for a value of zero or one of 50%, each. To ensure the hazard rates to be well-behaved for all possible covariate combinations, the normally distributed covariate x_2 was restricted to be larger than zero and smaller than 10, but restricting the covariate should not have a relevant effect on the results, as values outside the predefined range are expected to occur very rarely.
The Cox-type subdistribution hazards regression model was set to be

$$\begin{aligned}\gamma_1(t|\mathbf{x}) &= \gamma_{1;0}(t)\, \exp(\boldsymbol{\eta_1}^\top \mathbf{x}) \\ &= 0.001 \exp\left(-\frac{0.001\,t}{\ln(1.3)}\right) \exp\big(\ln(1.15)\,x_1 + \ln(0.9)\,x_2 + \ln(2)\,x_3\big).\end{aligned}$$

Choice of the model implies time-constant subdistribution hazard ratios for all three covariates considered of 1.15, 0.9, and 2. To fulfil the restrictions presented in Section 6.3.2, the cause-specific hazard for the event of interest had to be chosen appropriately, so that

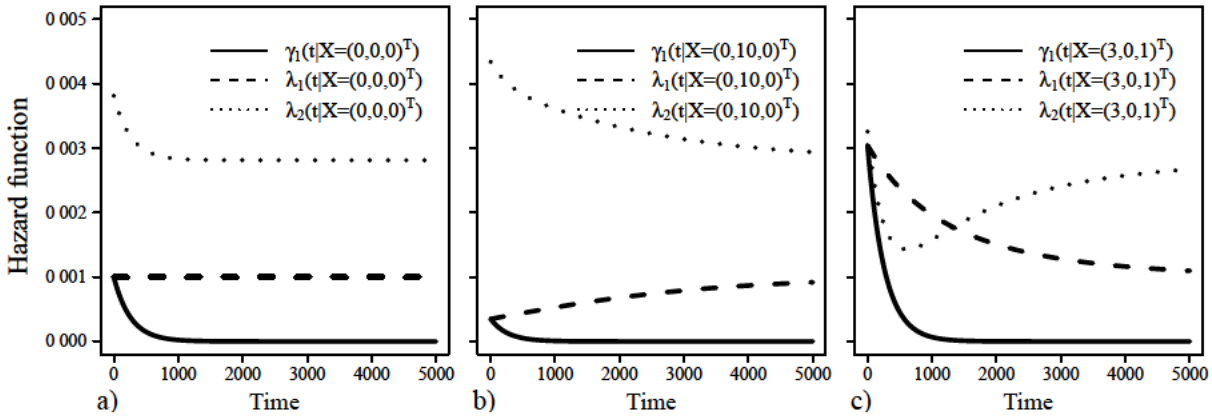

Figure 6.5: Example 5 – Cause-specific and subdistribution hazard functions for three individuals with different risk profiles for the multiple regression model. The solid line shows the subdistribution hazard for the event of interest, the dashed line the cause-specific hazard for the event of interest and the dotted line the cause-specific hazard for the competing event. Left: Individual with all covariates set to zero, i.e. $\mathbf{X}{=}(0,0,0)^T$. Middle: Individual with low subdistribution hazard for the event of interest, $\mathbf{X}{=}(0,10,0)^T$. Right: Individual with high subdistribution hazard, $\mathbf{X}{=}(3,0,1)^T$.

it converges to the value of the subdistribution hazard for a given $\mathbf{x}$ for t going to zero. Here, the model for the cause-specific hazard rate for k=1 was chosen to be

$$\lambda_1(t|\mathbf{x}) = 0.001 \exp\big(\ln(1.15)\, \exp(-0.0005\, t)\, x_1 + \\ + \ln(0.9)\, \exp(-0.0005\, t)\, x_2 + \ln(2)\, \exp(-0.0005\, t)\, x_3\big),$$

implying time-dependent cause-specific hazard ratios. The cause-specific hazard rate for the competing event $\lambda_2(t|\mathbf{x})$, which is necessary for the simulation algorithm to obtain the desired subdistribution hazards model, was computed following Equation 6.7. The subdistribution hazard rates for the event of interest and the cause-specific hazard rates for both types of event are illustrated in Figure 6.5 for the baseline case of $\mathbf{X}$=$\mathbf{0}$ (Figure a), for an individual with low subdistribution hazard for an event of interest, having covariate values of $\mathbf{X}$=$(0, 10, 0)^\top$ (b), and for an individual with a high subdistribution hazard, having a covariate vector of $\mathbf{X}$=$(3, 0, 1)^\top$ (c).
Generation of competing risks data was repeated 4,000 times with 1,000 subjects, each. Data were analysed using the function *crr* in the library *cmprsk* of the statistical software R to obtain estimates of the regression coefficients for a proportional subdistribution hazards regression model. Results of the simulations are summarized in Table 6.5, bold numbers denote the desired regression coefficients. Coverage proportions of 95% confidence intervals were derived as described for Example 2. The R code used for generation of competing risks data for the multiple subdistribution hazards regression example is presented in the Appendix (Section A).

Covariate	$\mathbf{X_1 \sim U(0, 3)}$	$\mathbf{X_2 \sim N(5, 1)}$	$\mathbf{X_3 \sim B(1, 0.5)}$
$\boldsymbol{\eta_1}$	$\mathbf{\ln(1.15)=0.140}$	$\mathbf{\ln(0.9)=-0.105}$	$\mathbf{\ln(2)=0.693}$
Mean	0.140	-0.107	0.698
Median	0.140	-0.108	0.695
Std.dev.	0.073	0.065	0.134
$Q_{.025}$	0.002	-0.236	0.442
$Q_{.975}$	0.281	0.023	0.964
CI coverage	0.960	0.943	0.953

Table 6.5: Example 5 – Summaries of estimated regression coefficients from multiple proportional subdistribution hazards regression models based on 4,000 generated datasets with 1,000 observations, each. True values of subdistribution log-hazard ratios are shown in bold font.

6.4 Discussion on simulation of competing risks data

Different methods for simulation of competing risks data were used in the literature in order to evaluate and compare statistical methods for analysis of event time data in the presence of multiple types of event. Beyersmann et al (2009) recommended to use the cause-specific hazards for simulation of competing risks data, as cause-specific hazards "completely determine the competing risks process", and presented an algorithm for generation of competing risks data following prespecified cause-specific hazards.

Due to the growing interest in the subdistribution hazard, which is directly linked to the cumulative incidence function, different methods focussing on that quantity were introduced in recent years (e.g Katsahian et al, 2006; Ruan and Gray, 2008; Sun et al, 2008). In order to evaluate or compare these methods and to find the best available procedure for a given data situation, competing risks data providing a prespecified subdistribution hazard have to be generated. In many research articles competing risks data with predefined subdistribution hazards are simulated from a unit exponential mixture distribution (Fine and Gray, 1999). Beyersmann et al (2009) presented how cause-specific hazards can be chosen to obtain competing risks data following the desired subdistribution hazards. As this will generally lead to time-dependent overall hazards, which are considered for generation of event times, appropriate methods for even-time simulation are inevitable. Generation of event time data following flexible time-varying hazard rates can be performed using the inversion method commonly considered for data simulation. An alternative approach based on the Binomial Algorithm, which was provided by Sylvestre and Abrahamowicz (2008) for simulation of time-to-event data in the presence of time-varying covariates, is presented. In complex scenarios, requiring numerical approximation of the cumulative cause-specific hazard rates and consequently of the cumulative overall hazard rate, and numerical methods for determination of the generated event time from the cumulative overall hazard rate and a uniform random variable, the inversion method can become computationally very expensive. The approach based on the Binomial Algorithm generates event times in discrete time, which might lead to bindings in the simulated event times. The number of bindings can be kept low by choosing small baseline hazard rates, which will increase the computation time, as more timepoints have to be considered for each individual. So, hazard rates have to be chosen carefully, in order to obtain adequate event times and to limit computational burden.
The data generating process using the Binomial Algorithm for simulation of competing risks data with predefined subdistribution hazard rates was validated for different scenarios including time-dependent hazard ratios and multiple regression models. For all examples a good behaviour of the data generating process was revealed.

Chapter 7

Simulation study

A simulation study was conducted to evaluate the performance of the mixture model approach for estimation of cause-specific and subdistribution hazards and consequently hazard ratios as presented in Section 4.5, using different settings for given cause-specific or subdistribution hazard rates, respectively. The spline-based approach, which was proposed in Section 5, with different values for the smoothing parameter μ was compared to parametric mixture model approaches assuming the conditional event times to follow Weibull or generalized gamma distributions (see Section 4.4.2). Competing risks data were generated for the following scenarios:

I) Time-constant cause-specific hazard ratio

II) Time-dependent monotonous cause-specific hazard ratio

III) Time-dependent non-monotonous cause-specific hazard ratio

IV) Time-constant subdistribution hazard ratio

Simulation of competing risks data is described in Section 7.1, the methods used for analysis of the simulated datasets are presented in Section 7.2. A detailed description of the scenarios considered can be found in Section 7.3. The main results of the simulation study, namely summaries of estimated hazard ratios for the prespecified hazard type, which is the cause-specific hazard ratio for Scenarios I to III and the subdistribution hazard ratio for Scenario IV, are presented in Section 7.4. Summaries for the estimated cause-specific hazards rates and subdistribution hazard rates as well as for the estimated hazard ratio, which was not presented in Section 7.4, are displayed in the Appendix (Section B.1). The simulation study was performed using the statistical software R (R Development Core Team, 2011). Sketches of the R code used for data generation and analysis of the simulated data can be found in the Appendix (Section B.2). Results of the simulation study are summarized and discussed in Section 7.5, including a table, that shows numbers and proportions of datasets with adequately derived maximum likelihood estimates (Table 7.7).
Results of the simulation study comparing the generalized gamma approaches and the spline approach with smoothing parameters of μ=1 and μ=100 for Scenarios II to IV with moderate censoring proportions (40% to 60% censored observations) were also described in a manuscript that was submitted for publication and was under review when this work was finalized. The article was written in collaboration with Prof. Dr. Georg Schmidt from the

1. Medizinische Klinik of the Klinkum rechts der Isar (Technische Universität München) and Prof. Dr. Kurt Ulm of the Institut für Medizinische Statistik und Epidemiologie of the Technische Universität München.

7.1 Generation of competing risks data

Simulation of competing risks data was performed as described in Section 6. Data generation for a given cause-specific hazard ratio (Scenarios I to III) was conducted using the method proposed by Beyersmann et al (2009), which is described in Section 6.2. When data were simulated in order to provide predefined subdistribution hazards (Scenario IV), the simulation algorithm based on the ideas of Beyersmann et al (2009), choosing the subdistribution hazard for the event of interest and one cause-specific hazard and determining the cause-specific hazard for the other event type, using the relationship between cause-specific and subdistribution hazards, as described in Section 6.3, was applied. The inversion method described in Bender et al (2005) and presented in Section 6.1 was used to draw event time data from a survival time distribution with the derived overall hazard rate, and event types were determined from Bernoulli experiments with probabilities proportional to the cause-specific hazard rates at the generated times of event.
For convenience and in correspondence to the method description in Section 5, data were generated for two independent groups without further consideration of covariates, and each individual could fail from one out of two possible types of event, the event of interest (k=1) and a competing event (k=2). The groups are called control group (X=0) and study group (X=1) throughout the section. For each simulation run 1,000 observations were generated. Subjects were randomly allocated to the control or the study group with a probability of 50%, each. Different censoring proportions were considered in order to investigate the effect of censored observations and limited information available on the precision of the estimates and the numerical stability of the estimating procedures:

a) Low censoring: Administrative censoring after 5 years leading to an amount of censored observations of 5-20%.

b) Moderate censoring: Administrative censoring after 5 years and additional drop-outs generated from a Weibull distribution with parameters $\lambda=\frac{1}{3}$ and σ=0.7, leading to proportions of censored observations of 40-60%.

c) High censoring: Administrative censoring after 5 years and additional drop-outs generated from a Weibull distribution with parameters λ=10 and σ=0.1, leading to proportions of censored observations of 70-80%.

Censoring times were generated independently and for each individual the observed time was calculated as the minimum of the generated event time, the potential drop-out time and the administrative censoring time of 5 years. The individual's status was chosen accordingly to be either the generated type of event, if the event time was smaller than the drop-out time and the administrative censoring time, or zero else, indicating a censored observation. For each scenario 500 simulation runs were performed.

7.2 Analysis of simulated data sets

For the spline based approach introduced in Section 5, data were analysed using penalized B-spline basis functions for estimation of conditional hazard rates in a mixture model as shown in Equation 5.7, using five interior knots located at the quantiles of the observed event times irrespective of the observed type of event and the covariate information. The same knot locations were used for both types of event. Lower slack knots were defined by the interval between timepoint zero and the first inner knot, upper slack knots were defined by the distance between the last inner knot and the maximum observed event time. Three different values for the smoothing parameter μ were investigated (μ=0.01, μ=1, and μ=100), leading to estimated conditional hazard rates that vary from rough, variable curves to smooth curves.
Three different parametric mixture models (Section 4.4.2) were investigated and compared to the spline models. A Weibull mixture model, assessing the group influence on parameter λ via $\lambda = \exp(\beta x)$, and two different generalized gamma mixture models. In a first generalized gamma model, the location parameter λ was allowed to vary between both groups as shown in Equation 4.26 and shape and scale parameters of the conditional event time distributions were assumed to be equal for control group and study group. This approach is later also called GG_λ-approach. In a saturated generalized gamma mixture model, later also referred to as $GG_{\lambda\tilde{\alpha}\nu}$, group effects on all three parameters for each type of event were assessed, as described in Section 2.4.3 for the standard survival setting.
Parameter estimation for all scenarios was performed by numerical maximization of the log-likelihood functions shown in Equations 4.24, 4.25 or 5.11, respectively, using the R-function *nlm*. In order to derive adequate starting values for the parametric mixture models, an exponential mixture model (see Equation 4.22) was fit to the data in a first step using starting values of zero for the regression coefficients of the marginal event type distribution and for the regression coefficients indicating group differences in the conditional event time distributions. As starting values for the regression coefficients describing the conditional baseline log-hazard rates, i.e. the logarithms of the conditional hazard rates for the control group, ln(0.1) was chosen. Results for the exponential mixture model are not shown, as the exponential model implies time-constant conditional hazard rates and therefore does not allow flexible estimation of cause-specific and subdistribution hazard rates. For the Weibull model the regression coefficients derived from the exponential model were used as starting values for the regression coefficients of the marginal event type distribution and the location parameters of the conditional event time distributions. A starting value of one was considered for the shape parameters of the conditional event time distributions. For the generalized gamma model assessing group effects on the location parameter only (GG_λ), the according estimates of the Weibull model (the estimate for α was transformed to $1/\hat{\alpha}$ as starting value for $\tilde{\alpha}$) and a value of one for parameters ν_1 and ν_2 were used as starting values. The parameters derived from that model were used as starting values for the saturated generalized gamma model and starting values for regression coefficients indicating group differences in shape and scale parameters of the conditional event time distributions were set to zero. For the spline-based approach the coefficients for the marginal event type distribution estimated for the exponential mixture model were used as starting values for those parameters, for the conditional event time parameters values of -0.5 were used for the baseline parameters (X=0) and values of zero for the parameters assessing

differences in the conditional hazard rates between the study group and the control group. Estimates for cause-specific and subdistribution hazard rates and consequently hazard ratios were derived from the mixture models as described in Section 4.5. For a sequence of 500 equidistant timepoints summaries of estimated hazard rates and log-hazard ratios were calculated (mean, median, 5% and 95% quantiles). These summaries of estimated cause-specific and subdistribution hazard rates and hazard ratios are displayed graphically in Section 7.4 or in the Appendix (Section B.1). Additionally, for the scenarios with true time-constant hazard ratios (Scenarios I and IV) summaries of the average cause-specific or subdistribution log-hazard ratios, respectively, are presented. Average log-hazard ratios were derived as means of estimated log-hazard ratios for timepoints with an observed event of interest. For better interpretability and comparison to the predefined hazard ratios, summary statistics of estimated log-hazard ratios for given timepoints and for average log-hazard ratios were back-transformed to the scale of hazard ratios. While medians and quantiles of the exponentiated log-hazard ratios equal medians and quantiles of the hazard ratios, exponentiated arithmetic means of the log-hazard ratios equal the geometric means of the hazard ratios. This is indicated in the according figures and tables, but not further mentioned and discussed in the text. Variance and mean squared error (MSE), calculated as squared bias plus variance, are presented for average log-hazard ratios.
Maximum likelihood estimates for the mixture model regression coefficients could not be derived adequately for all simulated datasets for the generalized gamma models and the spline approach with a smoothing parameter of μ=0.01, as the maximization algorithm did not converge. Numerical problems for estimation of parameters in generalized gamma models were observed and discussed before (Gomes et al, 2008; Noufaily and Jones, 2013). Numbers of datasets without adequately derived estimates are presented in the text and are summarized in Table 7.7. Datasets, for which maximum likelihood estimates could not be obtained appropriately, were removed from analyses of model performance for the according model. Additionally, cause-specific or subdistribution log-hazard ratios of plus or minus infinity were derived for the generalized gamma mixture models for some timepoints in a small number of datasets. These datasets were also not considered for calculation of summary statistics.

7.3 Simulation scenarios

7.3.1 Scenario I - Constant cause-specific hazard ratio

In a first scenario a time-constant cause-specific hazard was considered for both types of event for both groups, leading to a constant cause-specific hazard ratio. The cause-specific hazards were chosen as follows:

$$\lambda_1(t|X{=}0) = \frac{1}{3} \approx 0.33$$
$$\lambda_1(t|X{=}1) = 0.25$$
$$\lambda_2(t|X{=}0) = 0.20$$
$$\lambda_2(t|X{=}1) = 0.20$$

So the true underlying cause-specific hazard ratio for the event of interest (k=1) is 0.75 ($HR^{cs}_{k=1}$=$\lambda_1(t|X{=}1)/\lambda_1(t|X{=}0)$), but no group effect on the competing event (k=2) exists

($HR^{cs}_{k=2}$=1). Administrative censoring after five years is considered and additional drop-outs were generated using censoring distributions as described before to obtain different proportions of censoring.
Data were generated as described in Section 6.2. For each individual an event time was simulated from an exponential distribution with the individual's overall hazard rate $\lambda_{ov.}(t|x) = \lambda_1(t|x)+\lambda_2(t|x)$. The type of event was determined by a Bernoulli experiment with the probabilities for an event of type 1 or 2 proportional to the cause-specific hazard rates.
Results of the simulations with time-constant cause-specific hazard ratio can be found in Section 7.4.1 and in Section B.1.1.

7.3.2 Scenario II - Time-dependent monotonous cause-specific hazard ratio

In a second scenario the ability of the mixture model approaches to detect time-varying cause-specific hazards, translating to a time-dependent cause-specific hazard ratio, was investigated. The cause-specific hazard for the study group (X=1) was chosen to be constant over time as in the previous scenario, but the cause-specific hazard for the control group was chosen to decrease non-linearly over time, translating to a cause-specific hazard ratio for the event of interest that is increasing over time. The cause-specific hazard rates, which are illustrated e.g. in Figure B.16 for the control group and in Figure B.17 for the study group, were chosen to be

$$\begin{aligned}
\lambda_1(t|X{=}0) &= 0.2\left(1+\frac{3}{\exp(t)}\right)\\
\lambda_1(t|X{=}1) &= 0.2\\
\lambda_2(t|X{=}0) &= 0.2\\
\lambda_2(t|X{=}1) &= 0.2
\end{aligned}$$

leading to a cause-specific hazard ratio for the event of interest of

$$HR^{cs}_{k=1} = \frac{1}{1+3/\exp(t)} = \frac{\exp(t)}{\exp(t)+3}.$$

Again, competing risks data were generated using the algorithm provided by Beyersmann et al (2009) shown in Section 6.2. For simulation of event times, the cumulative overall cause-specific hazard rate, which is time-dependent for individuals of the control group, is necessary in order to apply the the inversion method presented in Section 6.1. The cumulative overall cause-specific hazard rate for the control group is

$$\begin{aligned}
\Lambda_{ov.}(t|X{=}0) &= \Lambda_1(t|X{=}0) + \Lambda_2(t|X{=}0) =\\
&= 0.2\left(t - 3\exp(-t) + 3\right) + 0.2\,t.
\end{aligned}$$

For the study group the cause-specific hazard rates were specified to be constant over time, giving a time-constant overall hazard rate of $\lambda_{ov.}(t|X{=}1) = \lambda_1(t|X{=}1)+\lambda_2(t|X{=}1) = 0.40$ and a cumulative overall cause-specific hazard rate of $\Lambda_{ov.}(t|X{=}1) = 0.40\,t$. So event times for the study group were drawn from an exponential distribution with hazard rate 0.40.

Event types were generated based on Bernoulli experiments as described above. Censored observations were again introduced using different distributions for drop-out times leading to the desired censoring proportions.

7.3.3 Scenario III - Time-dependent non-monotonous cause-specific hazard ratio

In a third scenario the cause-specific hazard rates were specified to lead to a cause-specific hazard ratio for the event of interest that is time-dependent and non-monotonous. While for the control group the underlying cause-specific hazard rate for the event of interest was determined to be increasing for early timepoints and decreasing afterwards, a U-shaped cause-specific hazard for the event of interest, decreasing early and increasing later, was chosen for the study group, translating to a cause-specific hazard ratio for the event of interest decreasing for early points in time and increasing later. The cause-specific hazards for the competing event again were set to be constant over time and equal for both groups. The cause-specific hazard rates were chosen to be:

$$\begin{aligned}
\lambda_1(t|X{=}0) &= 0.25 + \frac{0.5\,t}{\exp(t)} \\
\lambda_1(t|X{=}1) &= 0.05\,t + \frac{0.25}{\exp(t)} \\
\lambda_2(t|X{=}0) &= 0.2 \\
\lambda_2(t|X{=}1) &= 0.2
\end{aligned}$$

Illustrations of the true underlying cause-specific hazard rates and the true hazard ratio can be found in the corresponding figures (e.g. Figures B.31, B.32 and 7.7), which are presented in Section B.1.3 of the Appendix and in Section 7.4.3. Figures showing summaries of estimated subdistribution hazard rates and hazard ratios including illustrations of the expected values can be found in Section B.1.3. For simulation of event times the inversion method was applied, considering the sum of the cumulative cause-specific hazard rates for both event types

$$\begin{aligned}
\Lambda_{ov.}(t|X{=}0) &= \Lambda_1(t|X{=}0) + \Lambda_2(t|X{=}0) = \\
&= 0.25\,\exp(-t)\left((t+2)\,\exp(t) - 2\,t - 2\right) + 0.2\,t \\
\Lambda_{ov.}(t|X{=}1) &= \Lambda_1(t|X{=}1) + \Lambda_2(t|X{=}1) = \\
&= 0.025\left((t^2+10) - 10\,\exp(-t)\right) + 0.2\,t.
\end{aligned}$$

Censoring times were generated as described before.

7.3.4 Scenario IV - Constant subdistribution hazard ratio

In a fourth scenario the special case of a predefined constant subdistribution hazard ratio, as also considered in Example 2 of Section 6.3.4, was investigated. The true underlying subdistribution hazards for the event of interest $\gamma_1(t|X{=}0)$ and $\gamma_1(t|X{=}1)$ and the cause-specific hazard rates for the competing event $\lambda_2(t|X{=}0)$ and $\lambda_2(t|X{=}1)$ were chosen under

consideration of the restrictions and requirements described in the Section 6.3.2.
The subdistribution hazards for the two groups were chosen to be

$$\gamma_1(t|X{=}0) = 0.2 \exp(-0.25\,t)$$
$$\gamma_1(t|X{=}1) = 0.15 \exp(-0.25\,t)$$

translating to a time-constant subdistribution hazard ratio for the event of interest of $HR^{sd}_{k=1} = 0.75$.
The cause-specific hazards for the competing event were set to

$$\lambda_2(t|X{=}0) = 0.3 \exp(-0.1\,t)$$
$$\lambda_2(t|X{=}1) = 0.15$$

implying a time-dependent cause-specific hazard rate for the competing event for the control group and a time-constant hazard rate for the study group.
The cumulative subdistribution hazard rates for the event of interest are

$$\Gamma_1(t|X{=}0) = 0.8 - 0.8 \exp(-0.25\,t)$$
$$\Gamma_1(t|X{=}1) = 0.6 - 0.6 \exp(-0.25\,t)$$

and the cumulative cause-specific hazards for the competing event are

$$\Lambda_2(t|X{=}0) = 3 - 3 \exp(-0.1\,t)$$
$$\Lambda_2(t|X{=}1) = 0.15\,t.$$

The cause-specific hazard rates for the event of interest, which are needed for simulation of competing risks data providing the predefined subdistribution hazard rates, were derived numerically considering the relationship between cause-specific and subdistribution hazards shown in Equations 3.14 and 6.8. Event types were generated by Bernoulli experiments as described above and censoring times were introduced by administrative censoring and by simulated losses to follow-up as presented in Section 7.1.
Results are presented graphically and by summary statistics for the average subdistribution (log-)hazard ratio in Section 7.4.4 and in the Appendix in Section B.1.4.

7.4 Results of the simulation study

Results for the four simulation scenarios with three different censoring distributions, each, are shown in this section. Summaries of the estimated cause-specific hazard ratios (Scenarios I to III) or subdistribution hazard ratios (Scenario IV), derived as described in Section 7.2, are shown. For the scenarios with time-constant constant hazard ratios (I and IV) summaries of the average log-hazard ratios are presented.
Illustrations of the derived cause-specific and the subdistribution hazard rates as well as cause-specific hazard ratios (Scenario IV) and subdistribution hazard ratios (Scenarios I to III) can be found in the Appendix (Section B.1, Figures B.1 to B.60).

7.4.1 Constant cause-specific hazard ratio

In the first scenario a constant cause-specific hazard ratio was used. In the following the results for the different amounts of censoring are shown.

Low censoring

The results of Scenario I with a low proportion of censored observations are presented. Amount of censoring was between 5.8% and 11.4% with a mean censoring proportion of 8.8%. In Figure 7.1 summaries of the estimated cause-specific hazard ratios for the event of interest (k=1) are shown for the six models under consideration. Summaries of estimated cause-specific and subdistribution hazard rates and subdistribution hazard ratios can be found in Figures B.1 to B.5.
Maximum likelihood estimates could be derived for all datasets for all of the six models under consideration. In Table 7.1 summaries of the average cause-specific (log-)hazard ratios obtained from the different models are shown. Means, medians and quantiles were back-transformed as described in Section 7.2. Regarding the average hazard ratios, all investigated models performed similarly with comparable medians, means and variances leading to almost identical mean squared errors. According to Figure 7.1 and Table 7.1, means and medians of estimated hazard ratios are slightly lower than the underlying cause-specific hazard ratio, but very close to the true value of 0.75 for all models. True values of cause-specific and subdistribution hazard rates are detected well, as shown in the Appendix. The variability of obtained estimates, displayed by 5% and 95% quantiles in Figure 7.1, was smallest for the Weibull model and the generalized gamma model with equal shape

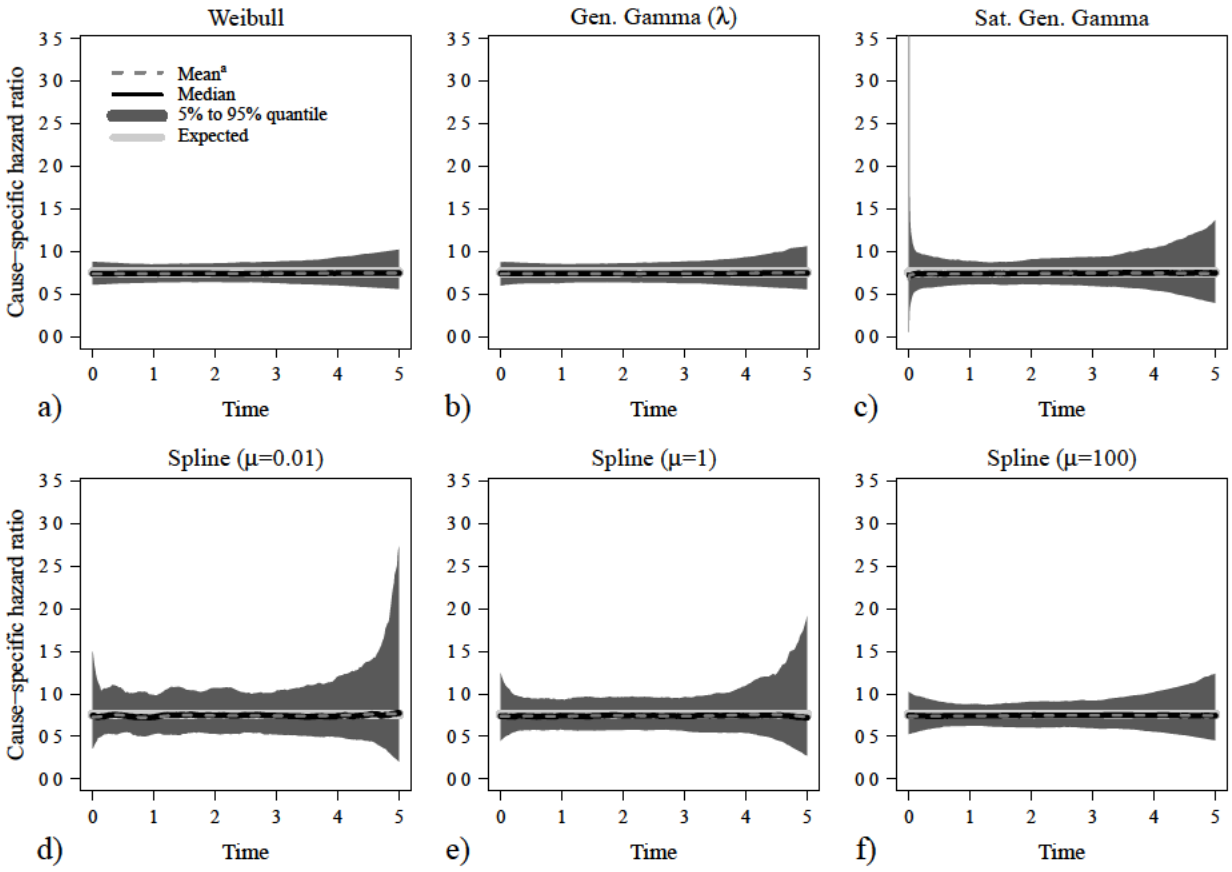

Figure 7.1: Scenario I - low censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

	Mean[a]	Median	$Q_{.05}$	$Q_{.95}$	Var (log)	MSE (log)
Weibull	0.743	0.747	0.644	0.857	0.008	0.008
GG_λ	0.743	0.747	0.644	0.856	0.008	0.008
$GG_{\lambda\tilde{\alpha}\nu}$	0.743	0.746	0.642	0.857	0.008	0.008
μ=0.01	0.742	0.745	0.642	0.856	0.008	0.008
μ=1	0.742	0.745	0.642	0.857	0.008	0.008
μ=100	0.743	0.746	0.642	0.859	0.008	0.008

Table 7.1: Scenario I - low censoring: Summary of estimated average cause-specific (log-)hazard ratios.
[a]Means of estimates for the average hazard ratio are exponentiated means of average log-hazard ratio estimates.

and scale parameters assumed for the conditional event time distributions for both groups. The spline based model with low penalization for roughness of the estimated conditional hazard rates (μ=0.01) provided the most variable results.

Moderate censoring

Results for the scenario with competing risks data following time-constant cause-specific hazard rates for both groups, but with a moderate amount of censoring, which was introduced by consideration of additional drop-outs as described in Section 7.1 leading to proportions of censored observations from 40.0% to 49.8% with a mean proportion of 46.2%, are presented here. Maximum likelihood estimates for the mixture model regression coefficients could be derived adequately for all datasets using the Weibull mixture model and the mixture models relying on the spline approach. For the generalized gamma mixture model assessing group differences for the location parameter only (GG_λ), the maximization algorithm converged for 498 datasets (99.6%), for the saturated generalized gamma model ($GG_{\lambda\tilde{\alpha}\nu}$) for 484 datasets (96.8%).
As was to be expected, estimates for the average cause-specific log-hazard ratios, which are summarized in Table 7.2, were more variable in the scenario with moderate amount of censoring compared to the results obtained in the presence of a lower amount of censoring, which were presented before. The true cause-specific log-hazard ratio of -0.288, translating to a hazard ratio of 0.75, was slightly underestimated by all methods. Variance and mean-squared errors were very similar for all models with a slightly higher MSE for the spline

	Mean[a]	Median	$Q_{.05}$	$Q_{.95}$	Var (log)	MSE (log)
Weibull	0.742	0.743	0.615	0.893	0.013	0.013
GG_λ	0.742	0.740	0.614	0.897	0.013	0.013
$GG_{\lambda\tilde{\alpha}\nu}$	0.740	0.739	0.612	0.897	0.013	0.013
Spline (μ=0.01)	0.739	0.741	0.608	0.895	0.015	0.015
Spline (μ=1)	0.740	0.741	0.611	0.894	0.013	0.013
Spline (μ=100)	0.742	0.742	0.614	0.896	0.013	0.013

Table 7.2: Scenario I - moderate censoring: Summary of estimated average cause-specific (log-)hazard ratios.
[a]Means of estimates for the average hazard ratio are exponentiated means of average log-hazard ratio estimates.

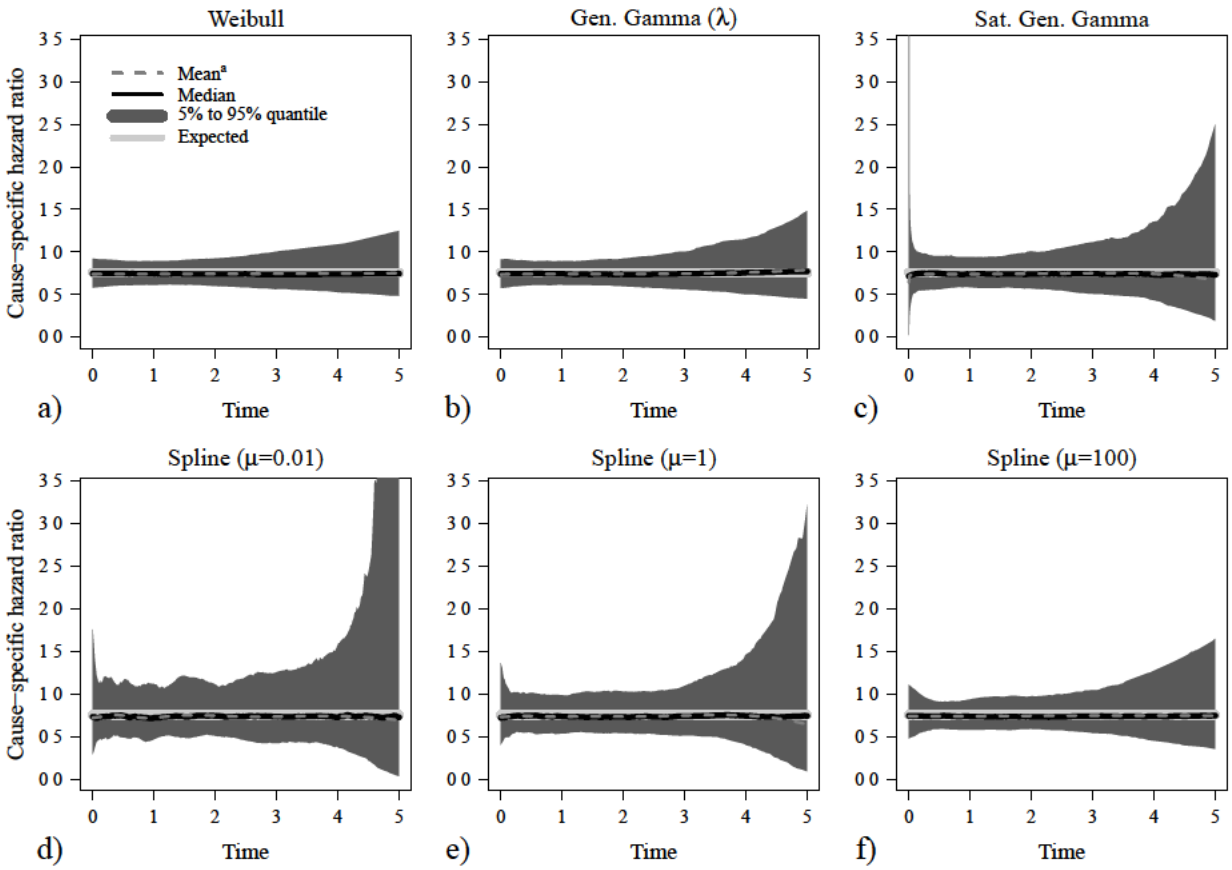

Figure 7.2: Scenario I - moderate censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantile at five years for the spline approach with μ=0.01 was 17.5.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

approach with the smoothing parameter set to μ=0.01. Summaries of cause-specific hazard estimates can be found in the Appendix in Figures B.6 and B.7, also indicating similar performance of all models, with the highest variability observed for the spline approach with μ=0.01, especially for late timepoints. This can also be seen in the summaries of the estimated cause-specific hazard ratios for the event of interest presented in Figure 7.2. According to Figures B.8, B.9 and B.10, which are shown in the Appendix, subdistribution hazard rates and hazard ratios were adequately derived from all mixture models.

High censoring

Using the third censoring distribution described in Section 7.1, censoring proportions between 68.9% and 77.6% were observed with a mean amount of censoring of 74.0%. The maximization algorithm for determination of maximum likelihood estimates did not converge for nine datasets using the generalized gamma approach (GG_λ) and for 81 datasets using the saturated generalized gamma approach (convergence proportions of 98.2% and 83.8%, respectively). Using the spline approach with a smoothing parameter of μ=0.01, maximum likelihood estimates could not be derived for one dataset (convergence proportion of 99.8%).

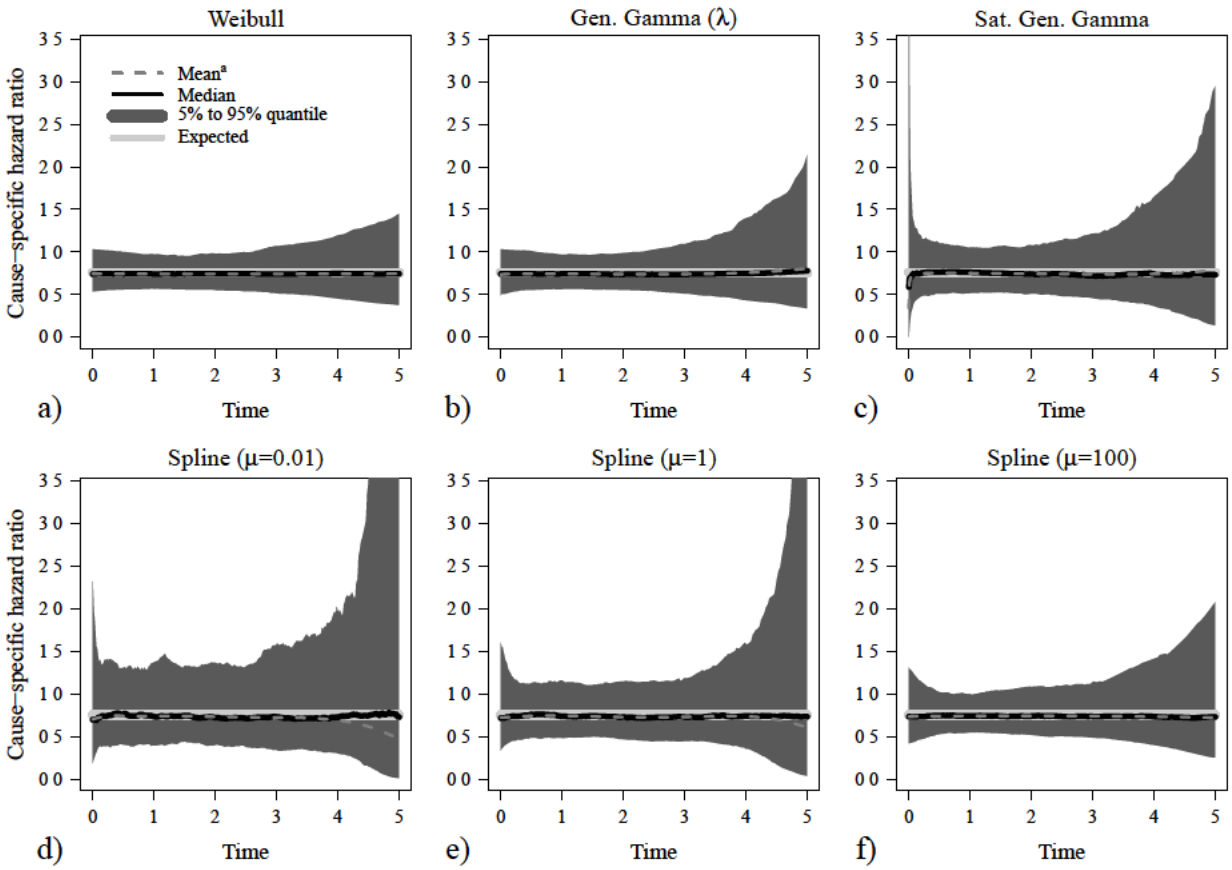

Figure 7.3: Scenario I - high censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantiles at five years not shown in the figures were 66.0 for the spline model with μ=0.01 and 7.4 for the spline model with μ=1.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Summaries of the estimated cause-specific hazard ratios are illustrated in Figure 7.3, confirming the results seen in Table 7.3. Results for the cause-specific hazard rates for both groups can be found in Figures B.11 and B.12, according summaries of the estimated subdistribution hazard rates and hazard ratios are displayed in Figures B.13 to B.15.
Estimates of average cause-specific (log-)hazard ratios are summarized in Table 7.3, showing similar means and medians for the different models, but slight differences in the variability of the estimates with the smallest variance and the smallest MSE observed for the Weibull model and the spline model with μ=100. For the spline-based approach with the smallest value for the smoothing parameter (μ=0.01) a high variance of estimates for the average cause-specific log-hazard ratio of 0.119, leading to the highest MSE of 0.120, was observed. The high variability was mainly caused by one dataset with an estimated average cause-specific log-hazard ratio of -7.01 (translating to an inverse cause-specific hazard ratio of 1,105), as the estimated cause-specific log-hazard ratio at the last timepoint with an observed event of interest ($t = 4.93$) was -284.3. Removing that dataset from the analysis resulted in a mean average log-hazard ratio for the spline approach with μ=0.01 of -0.299 (translating to a mean average hazard ratio of 0.742) and a variance of the average log-hazard ratios of 0.029 giving a mean-squared error of 0.029.

	Mean[a]	Median	$Q_{.05}$	$Q_{.95}$	Var (log)	MSE (log)
Weibull	0.743	0.743	0.563	0.959	0.026	0.026
GG_λ	0.743	0.743	0.563	0.976	0.027	0.027
$GG_{\lambda\tilde{\alpha}\nu}$	0.745	0.744	0.563	0.986	0.028	0.029
Spline (μ=0.01)	0.732	0.742	0.561	0.976	0.119	0.120
Spline (μ=1)	0.740	0.742	0.565	0.962	0.027	0.027
Spline (μ=100)	0.743	0.745	0.565	0.965	0.026	0.026

Table 7.3: Scenario I - high censoring: Summary of estimated average cause-specific (log-)hazard ratios.
[a]Means of estimates for the average hazard ratio are exponentiated means of average log-hazard ratio estimates.

7.4.2 Time-dependent cause-specific hazard ratio

The results of the simulations considering the second scenario with a true hazard ratio, that is increasing over time, resulting from one time-dependent and one time-constant cause-specific hazard rate, are presented here. The true underlying cause-specific hazard rates and the resulting hazard ratio for the event of interest are illustrated e.g. in Figures B.16, B.17 and 7.4 as solid grey lines. Graphical display of the estimated cause-specific hazard ratios can be found in this section, illustrations of estimates for cause-specific hazard rates and subdistribution hazard rates and hazard ratios are presented in the Appendix (Section B.1.2). As the true underlying cause-specific hazard ratio is time-dependent here, no estimates for the average log-hazard ratios were derived. The number and proportion of non-converged estimating procedures are mentioned in the text and are displayed in Table 7.7.

Low censoring

In a first setting the models were compared in the presence of a low amount of censoring (range of censored observations from 7.5% to 13.4% with a mean of 10.5%). In Figure B.16 estimated cause-specific hazard rates for the event of interest for the control group are shown, according results for the study group are displayed in Figure B.17. Summaries of the resulting cause-specific hazard ratios can be found in Figure 7.4.
With a low amount of censoring maximization algorithms converged for all investigated models for all 500 datasets, except for the saturated generalized gamma model, where maximum likelihood estimates could only be obtained for 492 datasets (98.4%).
Figures B.16 and B.17, showing summaries of estimates for the cause-specific hazard rate for the event of interest, reveal that for the Weibull model (Figure a), allowing only the location parameters λ_k of the conditional event time distributions to vary between groups and assuming the parameters α_k to be the same for both groups, the central tendency of the estimated cause-specific hazard rates for the event of interest differs from the true hazard rate for both groups. A similar result was obtained for the generalized gamma model allowing only the location parameter to vary between both groups (GG_λ, Figure b), as the different shapes of the cause-specific hazard rates for both groups imply different shapes of the conditional hazard rates in the mixture model. The saturated generalized gamma approach ($GG_{\lambda\tilde{\alpha}\nu}$, Figure c), which allows $\tilde{\alpha}_k$ and ν_k to vary between both groups, was

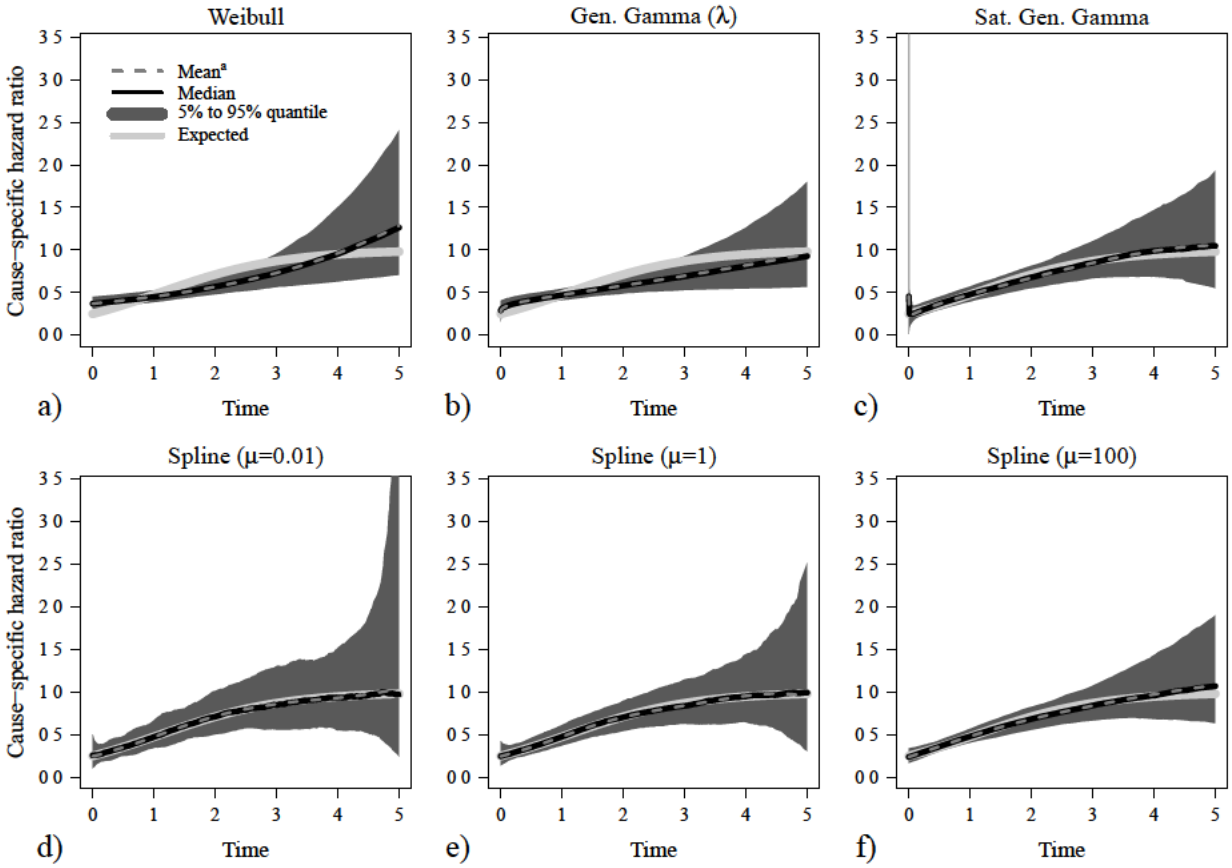

Figure 7.4: Scenario II - low censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantile at five years for the spline model with μ=0.01 was 4.7.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

able to reflect the true underlying cause-specific hazard rates for both groups well. For the saturated generalized gamma approach high variability and systematic bias was observed for very early timepoints. For the models using the spline approach to estimate conditional hazard rates in the mixture model with different values of the smoothing parameter (d-f), the true underlying cause-specific hazard rates and consequently the cause-specific hazard ratio were reflected well. Best results for the central tendencies (mean, median) were observed for the models with low values for the smoothing parameter (μ=0.01, μ=1), but at the cost of a higher variability compared to the model with smoother estimates of the conditional hazard rates (μ=100). This was also observed for the subdistribution hazard rates and the subdistribution hazard ratio, for which summaries are displayed in Figures B.18, B.19 and B.20.

Moderate censoring

Summaries of the estimated cause-specific hazard ratios derived from the different mixture models can be seen in Figure 7.5. According results for the cause-specific hazard rates are displayed in Figures B.21 and B.22. Mean proportion of censored observations was 44.9% (range from 39.9% to 49.0%). For the Weibull approach the algorithm for numerical maximization of the log-likelihood, intended to find maximum likelihood estimates for the

regression coefficients, converged for all of the 500 generated datasets. The algorithm did not converge adequately for one of the 500 datasets for the generalized gamma approach assuming shape and scale parameters of the conditional event time distributions to be the same for both groups (convergence for 99.8% of the datasets). For the saturated generalized gamma approach maximum likelihood estimates could be derived for 465 (93.0%) of the datasets. When the spline approaches with different values for the smoothing parameter μ were used, the maximization algorithm converged for all datasets.
As in the setting with a low amount of censoring presented before, the parametric mixture models, assuming the conditional event times to follow Weibull distributions (Figure a) or generalized gamma distributions with same scale and shape parameters $\tilde{\alpha}_k$ and ν_k for both groups (b), could not reflect the true patterns of the cause-specific hazard rates and the true hazard ratio. The saturated generalized gamma model, allowing all parameters to vary between the groups (c), was capable to detect the the true underlying cause-specific and subdistribution hazard rates and ratios (see also Figures B.21 to B.25 in the Appendix), but again estimates for very early timepoints appear to be biased for the control group. Results for the spline approaches were similar to the setting with a low amount of censoring with the best fit for the central tendency observed for a small value of the smoothing parameter,

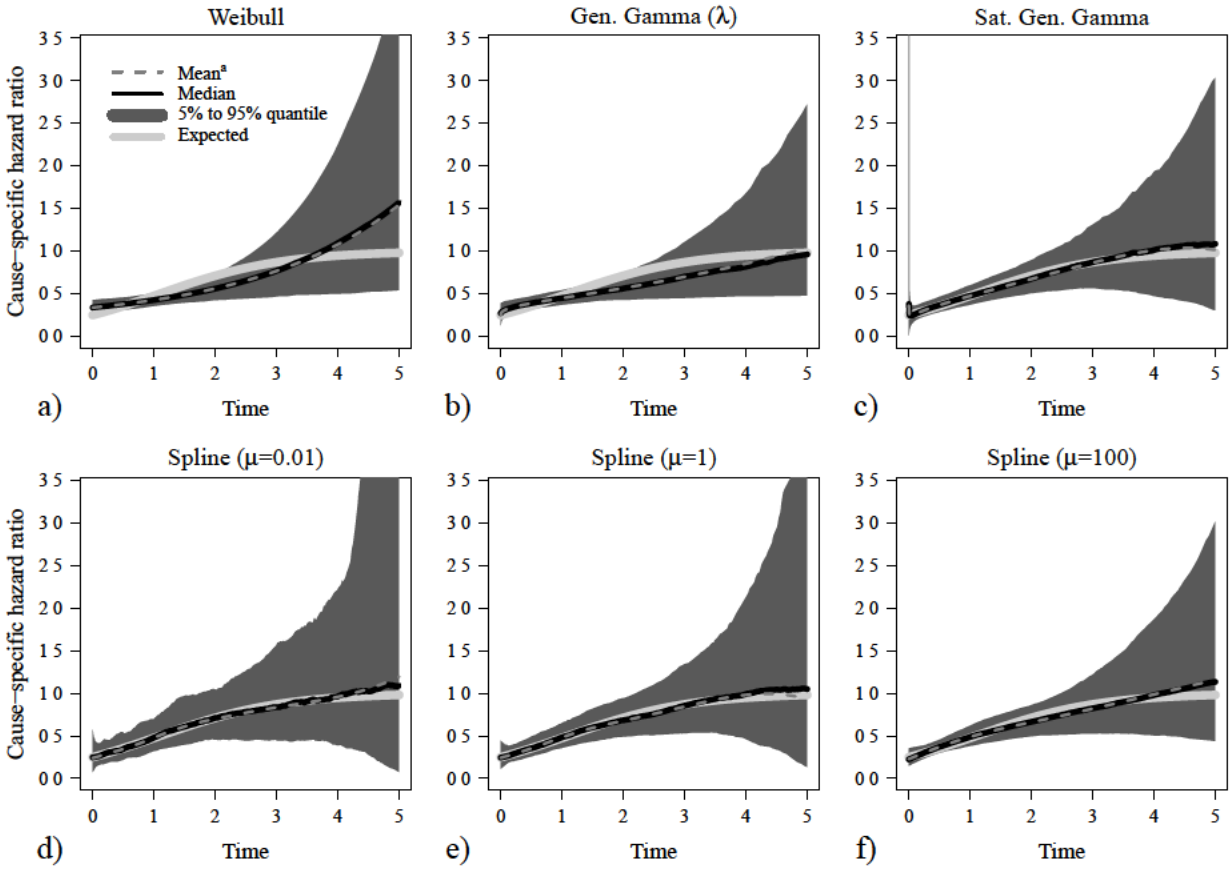

Figure 7.5: Scenario II - moderate censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantiles at five years not shown in the figures were 4.2 for the Weibull model 48.5 for the spline approach with μ=0.01 and 4.4 for the spline approach with μ=1.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

but lower variability for the models with a higher value of the smoothing parameter. As was to be expected, the higher amount of censored observations lead to a greater variability in the estimates represented by a wider distance from the 5% to the 95% quantile displayed in the figures. 95% quantiles for the cause-specific hazard ratio estimates at t=5, which were cut in Figure 7.5 in order to allow a better illustration of deviances of the central tendencies from the true cause-specific hazard ratio, were 4.2 for the Weibull model, 48.5 for the spline model with μ=0.01 and 4.4 for the spline approach with μ=1.

High censoring

Using the third censoring distribution described in Section 7.1, the proportion of censored observations ranged from 69.9% to 77.6% with a mean amount of censoring of 73.8%. For the generalized gamma model, investigating group effects on the location parameter only (GG_λ), the numerical maximization algorithm for estimation of the mixture model regression coefficients did not converge for 16 generated datasets, for the saturated generalized gamma model ($GG_{\lambda\tilde{\alpha}\nu}$) maximum likelihood estimates could not be determined for 116 generated datasets (proportion of converged algorithms of 96.8% and 76.8%). For the

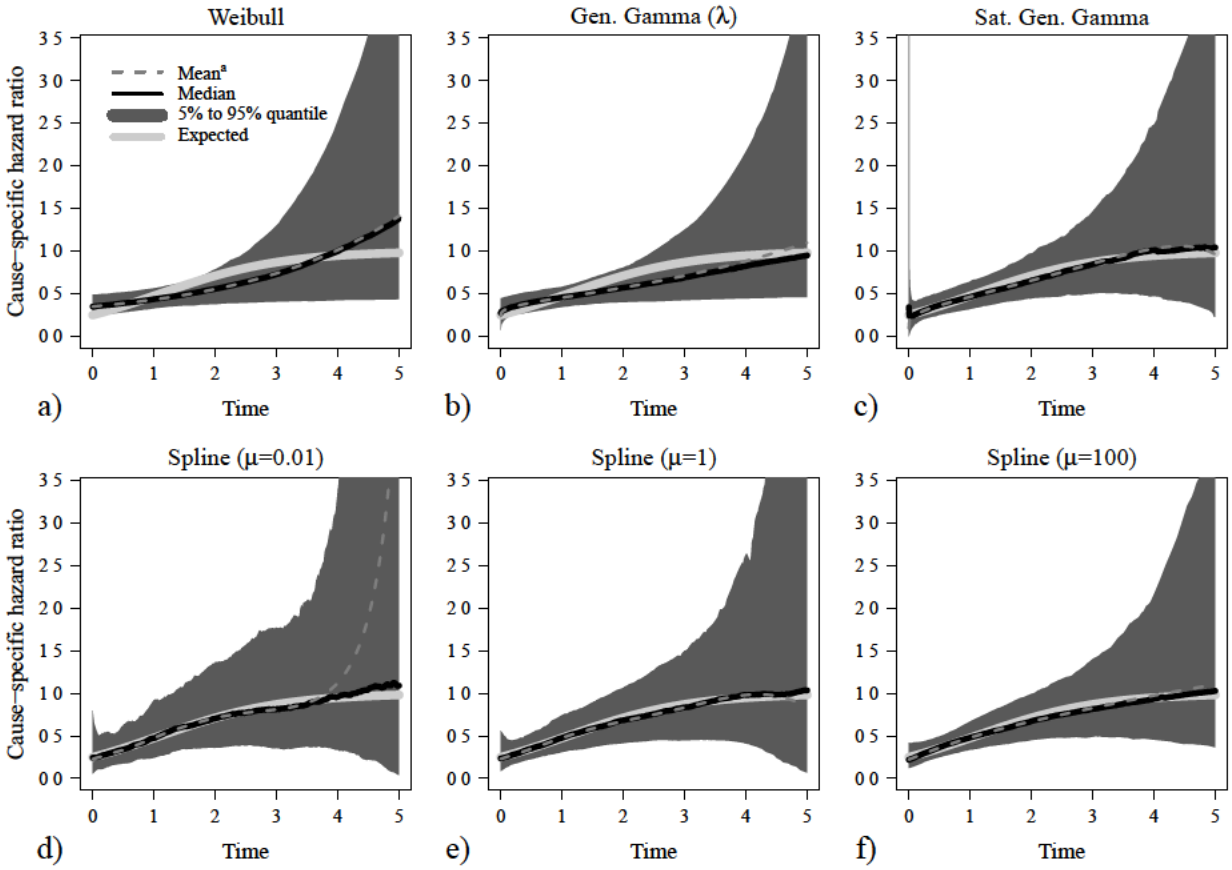

Figure 7.6: Scenario II - high censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantiles at five years not shown in the figures were 4.9 for the Weibull model, 4.2 for the GG_λ model, 4.7 for the $GG_{\lambda\tilde{\alpha}\nu}$ model, 2,930 for the spline model with μ=0.01, 8.2 for the spline model with μ=1, and 3.9 for the spline model with μ=100.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

spline approach with the lowest value for the smoothing parameter (μ=0.01) maximum likelihood estimates could be derived appropriately for 498 of the 500 datasets (99.6%). No numerical problems were observed for the Weibull model and the spline approaches with smoothing parameters of μ=1 and μ=100.
In Figure 7.6 results for the cause-specific hazard ratios are summarized for the investigated models, cause-specific hazard rates are shown in Figures B.26 and B.27, subdistribution hazard rates and hazard ratios are illustrated in Figures B.28 to B.30. As for the scenarios with lower censoring proportions, the spline approaches and the saturated generalized gamma model performed best regarding reflection of the true underlying cause-specific hazard ratio, but for the spline approach with a low value of the smoothing parameter large variability and a high mean was observed for estimates ot the cause-specific hazard ratio, especially for late timepoints. This was mainly caused by some extremely low estimates obtained for the cause-specific hazard rate of the control group (5% quantile at t=5: 2.4×10^{-5}), leading to very high estimates for the cause-specific hazard ratio (larger than 100 for 36 of the generated datasets at t=5). 95% quantiles of the estimated cause-specific hazard ratios at five years, which are all cut in the figures, were 4.9 for the Weibull mixture model, 4.2 for the GG_λ model and 4.7 for the saturated generalized gamma model. For the spline models the corresponding 95% quantiles were 2,930 for the spline model with μ=0.01, 8.2 for the spline model with μ=1, and 3.9 for the spline model with μ=100.

7.4.3 Non-monotonous cause-specific hazard ratio

For the third scenario, which is described in detail in Section 7.3.3, a non-monotonous cause-specific hazard ratio for the event of interest was considered, which is caused by a cause-specific hazard for the event of interest in the reference group, that is increasing for early timepoints and decreasing later, and a cause-specific hazard for the study group, that decreases first and increases for later points of time. An illustration of the hazard functions used for data generation can be found e.g. in Figures B.31 and B.32.

Low censoring

When data were generated using the cause-specific hazard rates as described in Section 7.3.3 and considering only administrative censoring after 5 years, but no additional dropouts, the censoring proportions observed in the simulation runs ranged from 8.3% to 13.7%, with a mean censoring proportion of 10.9%. With the low amount of censoring maximum likelihood estimates of the mixture model regression coefficients could be derived for all datasets for all models, except for the saturated generalized gamma model. Estimates for the mixture model regression coefficients were obtained for 454 datasets (90.8%) when that model was applied.
In Figure 7.7 summaries of the estimated cause-specific hazard ratios for the event of interest for all models under investigation are displayed. The figure reveals that the true shape of the cause-specific hazard ratio could not be reflected by the Weibull mixture model and the GG_λ approach, as these models do not allow for different shapes of the conditional event time distributions in both groups, which lead to biased estimates for the cause-specific hazard rates (Figures B.31 and B.32 a and b in the Appendix). Best results regarding the mean or the median of the estimated cause-specific hazard ratios

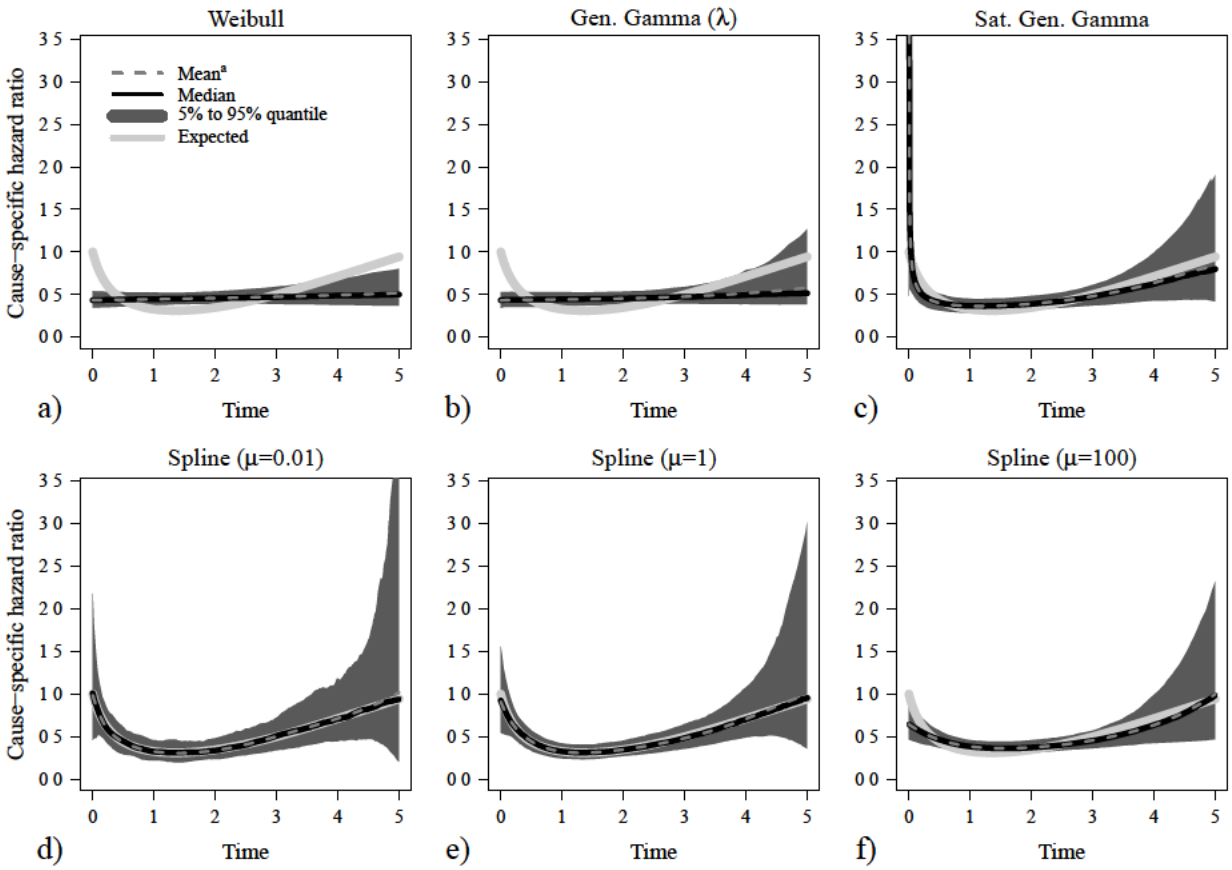

Figure 7.7: Scenario III - low censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantile at t=5 for the spline model with μ=0.01 was 4.8.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

were obtained using the spline approach with smoothing parameters of μ=0.01 or μ=1, but high variability was observed for late timepoints. For the spline approach with a smoothing parameter of μ=100 estimates for the conditional hazard rates for both groups tend towards a straight line due to higher penalization (Figures B.31 and B.32 f in the Appendix), resulting in less flexible estimates for the cause-specific hazard ratio (Figure 7.7 f). Mean and median of the estimated cause-specific hazard rates obtained from the saturated generalized gamma model are close to the true values, but small deviations from the true cause-specific hazard rates and consequently the true cause-specific hazard ratio are present (Figures B.31, B.32 and 7.7 c). Additionally, high variability for very early timepoints was observed for the $GG_{\lambda\tilde{\alpha}\nu}$ approach, especially in estimates for the study group (X=1). Similar results were found for the subdistribution hazard rates and the subdistribution hazard ratios, which were derived from the mixture models (results shown in Figures B.33 to B.35). Means and medians of the estimates were closest to the true subdistribution hazards using the spline approaches with smoothing parameters of μ=0.01 and μ=1. The spline approach with a high value of the smoothing parameter (μ=100) and the saturated generalized gamma mixture model lead to subdistribution hazard rates that were close to the true value, but showed small deviations for some timepoints. The Weibull and the GG_{λ} approach were not able to detect the true subdistribution hazard rates.

Moderate censoring

Using the second censoring distribution described in Section 7.1, censoring proportions of 42.3% to 52.2% with a mean of 47.7% were observed. Maximum likelihood estimates of mixture model regression coefficients could be determined for all datasets for the Weibull approach, for the generalized gamma approach allowing only the location parameters of the conditional event time distributions λ_k to vary between the groups, and for the three spline models (μ=0.01, μ=1, and μ=100). For the mixture model assuming the conditional event times to follow generalized gamma distributions allowing also shape and scale parameters ($\tilde{\alpha}_k$, ν_k) to differ between both groups, regression coefficients could be estimated adequately for 417 datasets (83.4%).
The estimated cause-specific hazard ratios are displayed in Figure 7.8. As was to be expected, the variability of the derived estimates was higher compared to the results presented in Figure 7.7 due to the higher proportion of censored observations and consequently the lower amount of information available for parameter estimation.
As before, estimates of the cause-specific hazard ratios derived from the spline-based approaches with small values for the smoothing parameter μ showed a large variability, especially for later timepoints with reduced risk sets, but mean and median of the estimated

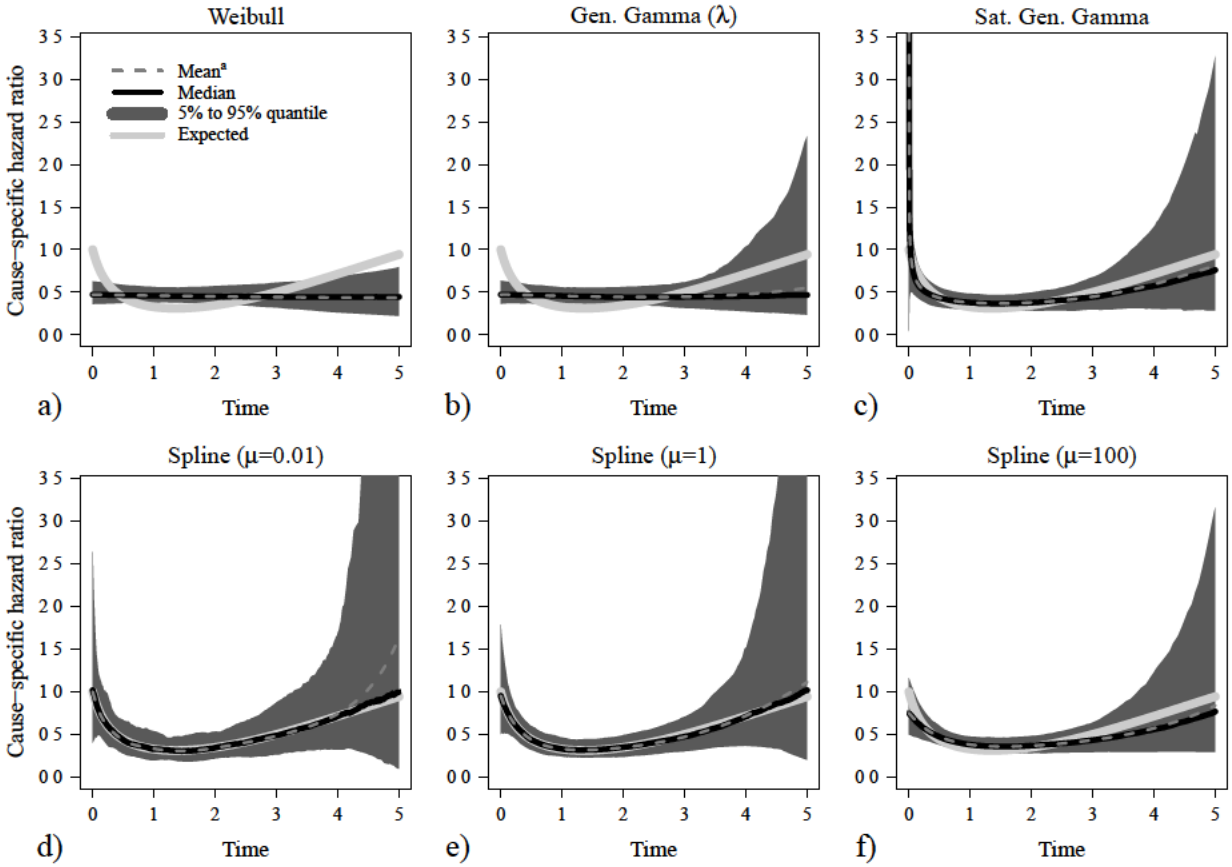

Figure 7.8: Scenario III - moderate censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantile for the spline methods with μ=0.01 and μ=1 were 128.7 and 9.7, repectively.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

cause-specific hazard ratios are near the true cause-specific hazard ratio for all timepoints. Due to estimates for the cause-specific hazard rates of the control group, that are very close to zero for late timepoints, when a smoothing parameter of μ=0.01 is used (see Figure B.36 a), some estimates for the cause-specific (log-)hazard ratio of the event of interest are very high for late timepoints and the mean of the estimated hazard ratios is larger than the true cause-specific hazard ratio. As the cause-specific hazard rates for the spline approach with a smoothing parameter of μ=100 tend towards a straight line due to the penalization of differences between nearby regression coefficients, estimates for the cause-specific hazards and consequently the hazard ratio were found to be slightly biased. The Weibull and the GG_λ approach do not reflect the true cause-specific hazard rates and hazard ratios well, for the saturated generalized gamma approach mean and median hazard rates are close to the true value, but small deviances from the true value are present. The 95% quantiles at 5 years, which are not shown in Figure 7.8, are 128.7 for the spline approach with μ=0.01, and 9.7 for the spline approach with μ=1. Results for the subdistribution hazards and the subdistribution hazard ratio can be found in the Appendix in Figures B.38, B.39 and B.40. Deviations from the expected subdistribution hazards are similar to those described for the cause-specific hazards.

High censoring

With the censoring distribution leading to a high amount of censored observations, censoring proportions of 70.5% to 78.1% with a mean proportion of censored observations of 74.6% were observed. Maximum likelihood estimates for the mixture model regression coefficients could no be derived adequately for the generalized gamma model for nine datasets (GG_λ, proportion of converged algorithms 98.2%), for 158 cases for the saturated generalized gamma model ($GG_{\lambda\tilde{\alpha}\nu}$, proportion of converged algorithms 68.4%) and for five datasets for the spline approach with the smallest value for the smoothing parameter (μ=0.01, proportion of converged algorithms 99.0%). For the spline approaches with μ=1 and μ=100 and for the Weibull mixture model maximum likelihood estimates could be determined for all 500 datasets.
Estimates for the cause-specific hazard ratios are summarized in Figure 7.9 and according results for the cause-specific hazards are displayed in the Appendix in Figures B.41 and B.42. Results for the subdistribution hazard rates and hazard ratio can be found in the Appendix in Figures B.43, B.44, and B.45. Again, the Weibull mixture model and the generalized gamma mixture model, not allowing shape and scale parameters to vary between both groups, could not reflect the true shape of the cause-specific and the subdistribution hazard rates. The (geometric) mean of cause-specific hazard ratio estimates for the spline approach with μ=0.01, derived as exponentiated mean of cause-specific log-hazard ratio estimates, also deviates substantially from the true cause-specific hazard ratio, mainly caused by estimates for the cause-specific hazard for the event of interest in the control group, that are very close to zero (see Figure B.41 d). The median of the estimated cause-specific hazard ratios was close to the true underlying cause-specific hazard ratio for all considered timepoints. Again, the estimated cause-specific hazard rates for the spline approach with the smoothing parameter set to μ=100 tend towards a straight line, leading to slightly biased estimates of the cause-specific hazard rates and consequently of the hazard ratios.

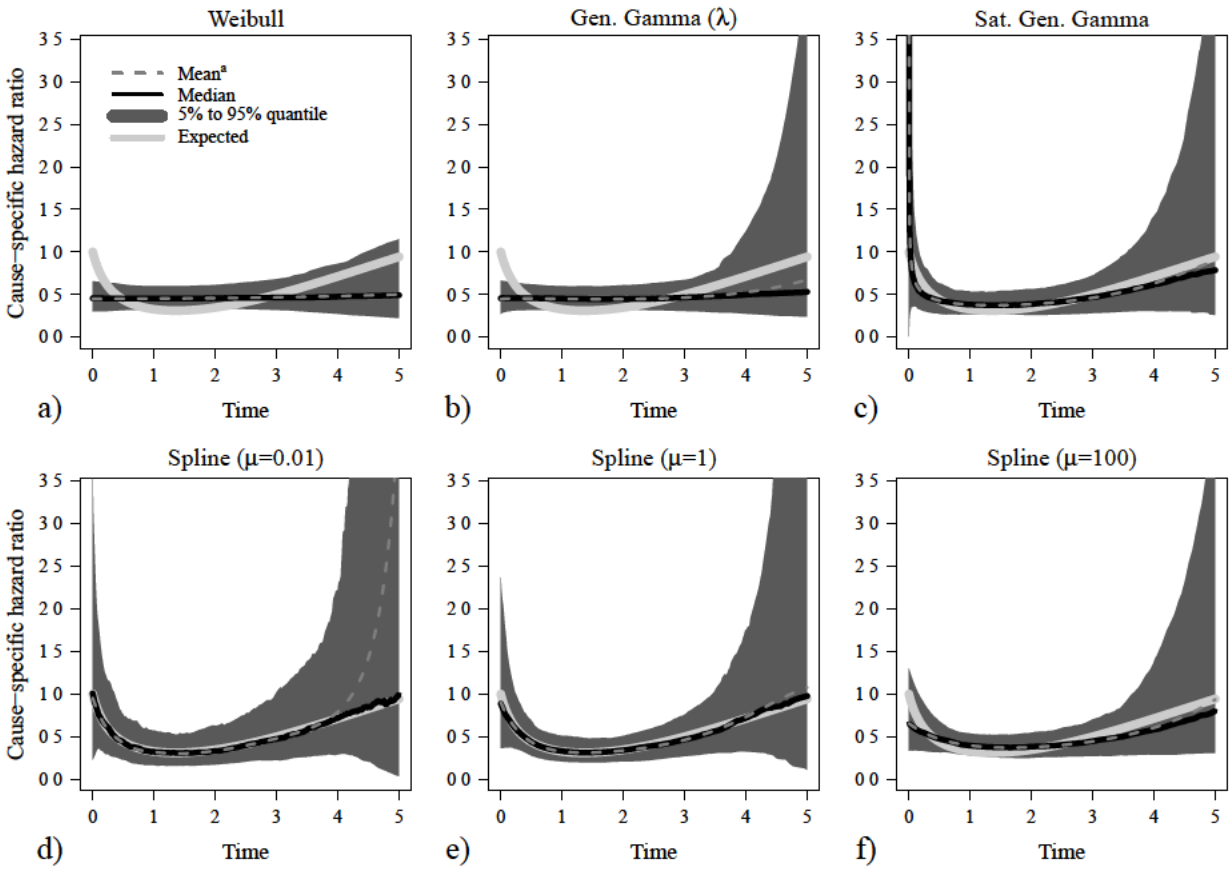

Figure 7.9: Scenario III - high censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods. 95% quantiles at t=5 not shown in the figures were 4.5 and 5.1 for the GG_λ and the $GG_{\lambda\tilde{\alpha}\nu}$ model, and 6,373, 13.2 and 4.4 for the spline models.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

7.4.4 Constant subdistribution hazard ratio

In this section the results of Scenario IV, representing the special case of a time-constant subdistribution hazard ratio as described in Section 7.3.4, are presented. As for that scenario the focus lies on the subdistribution hazard ratios, estimated subdistribution hazard ratios are summarized in the main body of the work. Results obtained for the cause-specific hazard ratios as well as illustrations of the estimated cause-specific and subdistribution hazard rates for both groups can be found in the Appendix (Section B.1.4).

Low censoring

In this setting, considering administrative censoring after five years of follow-up, but no additional drop-outs during follow-up, a mean censoring proportion of 16.0% was observed with censoring proportions in the generated datasets between 13.4% and 20.2%. Maximum likelihood estimates could be determined for all simulated data sets for the Weibull mixture model, the GG_λ mixture model, and the mixture models using the spline approach presented in Section 5. For the saturated generalized gamma model maximum likelihood estimates could not be determined for 15 datasets due to numerical problems.

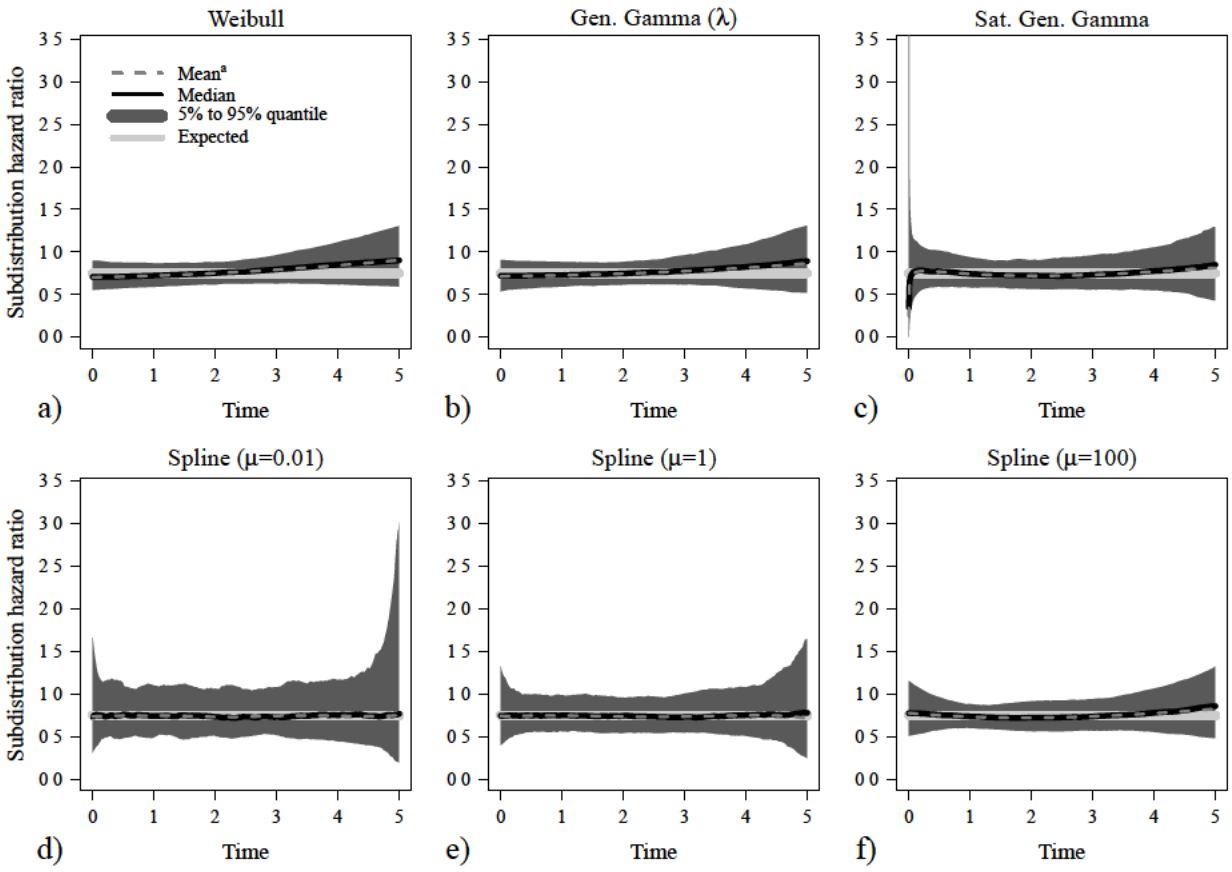

Figure 7.10: Scenario IV - low censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

An illustration of the true underlying subdistribution hazard for the event of interest for the control group can be found in Figure B.46 in the Appendix, represented as solid grey line, the corresponding quantity for the study group is displayed in Figure B.47. Results of the derived subdistribution hazard ratio estimates are summarized in Figure 7.10. Summaries for the estimated average subdistribution (log-)hazard ratios, determined as described in Section 7.2, are displayed in Table 7.4.

	Mean[a]	Median	$Q_{.05}$	$Q_{.95}$	Var (log)	MSE (log)
Weibull	0.753	0.755	0.627	0.883	0.011	0.011
GG_{λ}	0.749	0.751	0.626	0.879	0.011	0.011
$GG_{\lambda\tilde{\alpha}\nu}$	0.746	0.744	0.621	0.883	0.011	0.011
Spline (μ=0.01)	0.742	0.742	0.615	0.877	0.011	0.011
Spline (μ=1)	0.744	0.743	0.619	0.877	0.011	0.011
Spline (μ=100)	0.747	0.746	0.620	0.880	0.011	0.011

Table 7.4: Scenario IV - low censoring: Summary of estimated average subdistribution (log-)hazard ratios.
[a]Means of estimates for the average hazard ratio are exponentiated means of average log-hazard ratio estimates.

Figure 7.10 and Table 7.4 reveal similar properties for all investigated models with almost identical mean squared errors and no indication for systematic bias. Some deviations from the true cause-specific hazard rate of the control group were found for estimates derived from the parametric mixture models and the spline model with μ=100 for late timepoints, but not for the cause-specific hazard estimates for the study group (see Figures B.48 and B.49). For the saturated generalized gamma approach, variable estimates were observed for early timepoints again. For the spline approaches with small values of the smoothing parameter, higher variability was observed for late points of time.

Moderate censoring

When data were generated for the fourth scenario, as presented in Section 7.3.4, considering additional drop-outs using the censoring time distribution intended to produce a moderate amount of censored observations, the mean censoring proportion was 52.7.% with a minimum proportion of censored observations of 47.4% and a maximum proportion of 56.7%. Numerical maximization of the log-likelihood lead to valid results for all datasets using the Weibull mixture model, for 498 datasets (99.6%) for the GG_λ approach, and for 413 datasets (82.6%), when the saturated generalized gamma approach was used. No numerical problems occurred using the spline approaches.
Summaries of the average subdistribution (log-)hazard ratios are displayed in Table 7.5 and are illustrated in Figure 7.11. According results for the subdistribution hazard rates of both groups are shown in Figures B.51 and B.52 in the Appendix, estimates for the cause-specific hazard rates and the cause-specific hazard ratio can be found in Figures B.53, B.54, and B.55. Again all models performed similarly well with almost identical values for the mean-squared errors for the average subdistribution (log-)hazard ratios. When only those 413 datasets were considered, for which an adequate estimate could be obtained for the saturated generalized gamma approach, MSEs were 0.018 for all six models under investigation. As was to be expected, variances and consequently mean-squared errors were larger than in the setting with a lower amount of censored observations presented before. As observed before ,large variability was found for the estimated hazard ratios using the saturated generalized gamma approach for early timepoints and for the spline approach with a smoothing parameter of μ=0.01 for late timepoints.

	Mean[a]	Median	$Q_{.05}$	$Q_{.95}$	Var (log)	MSE (log)
Weibull	0.748	0.744	0.606	0.940	0.019	0.019
GG_λ	0.746	0.743	0.605	0.934	0.019	0.019
$GG_{\lambda\tilde{\alpha}\nu}$	0.745	0.746	0.597	0.927	0.018	0.018
Spline (μ=0.01)	0.741	0.735	0.597	0.929	0.019	0.019
Spline (μ=1)	0.744	0.740	0.600	0.929	0.019	0.019
Spline (μ=100)	0.747	0.743	0.602	0.935	0.019	0.019

Table 7.5: Scenario IV - moderate censoring: Summary of estimated average subdistribution (log-)hazard ratios.
[a]Means of estimates for the average hazard ratio are exponentiated means of average log-hazard ratio estimates.

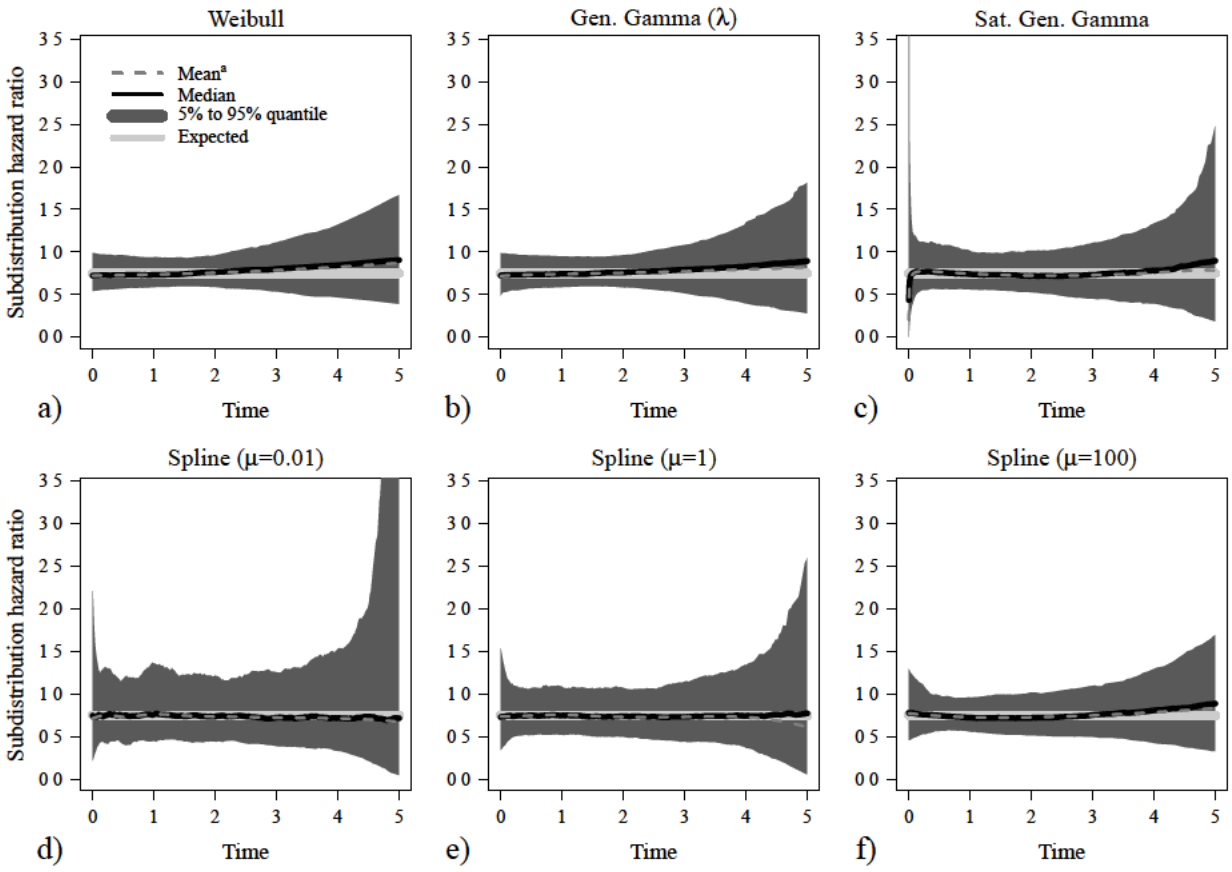

Figure 7.11: Scenario IV - moderate censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods. 95% quantiles at t=5 for the spline model with μ=0.01 was 8.9.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

High censoring

Variances of estimated cause-specific and subdistribution hazard ratios were higher than in the previous examples, when a censoring distribution leading to a higher amount of censored observations (72.6% to 80.0%, mean censoring proportion of 76.5%) was used, as was to be expected. Numerical maximization of the log-likelihood provided maximum likelihood estimates for all datasets, when the Weibull approach or the spline approach with smoothing parameters of μ=1 or μ=100 were applied. Maximum likelihood estimates could not be determined for 26 datasets for the GG_λ approach, for 159 datasets using the saturated generalized gamma approach, and for two datasets using the spline approach with μ=0.01 (converged algorithms for 94.8%, 68.2% and 99.6% of the datasets, respectively).
Summaries of the estimated average subdistribution (log-)hazard ratios for the event of interest can be found in Table 7.6. Graphical illustrations of the estimated subdistribution hazard ratios are given in Figure 7.12. As common for event time data with high amount of censored observations, the variability increases for later timepoints with less available information for all methods, with the highest variability observed for the spline approach with low penalization (μ=0.01). The 95% quantile at t=5 for the spline approach with

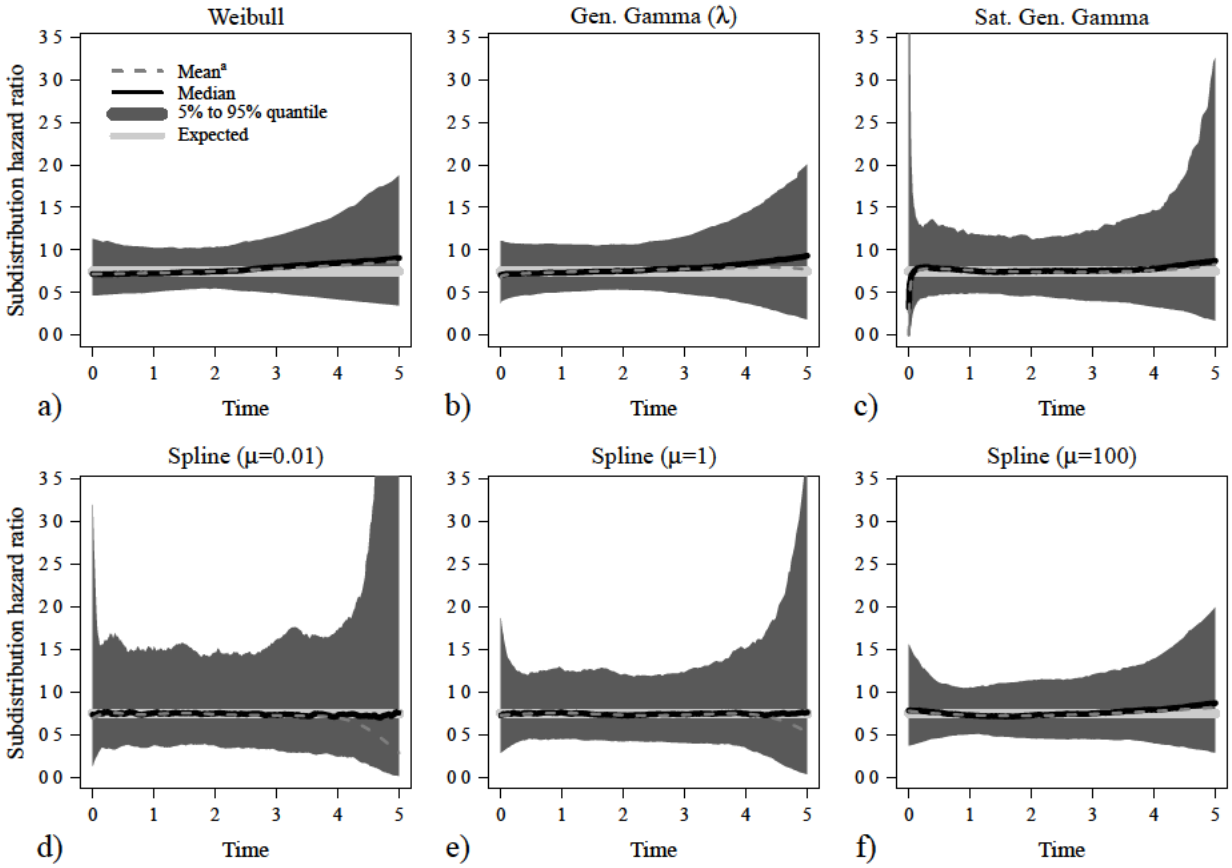

Figure 7.12: Scenario IV - high censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods. 95% quantiles at five years not shown in the figures were 31.3 for the spline model with $\mu = 0.01$ and 3.9 for the spline model with $\mu = 1$.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

μ=0.01, which cannot be seen in the figure, is 31.3, for the spline approach with μ=1 it is 3.9. Due to the high variability for late timepoints, the variance and consequently the mean-squared error of estimates for the average subdistribution log-hazard ratio obtained from the spline approach with μ=0.01 are larger than those for the other models, which provide comparable results (Table 7.6). The true cause-specific hazard rates showed similar deviations from the true values as described for the setting with low amount of censored observations (Figures B.58 and B.59). The high variability for late timepoints was also observed for the cause-specific hazard ratio, when the spline approach with the small value of the smoothing parameter was applied (Figure B.60).
When only datasets are considered, for which an adequate estimate for the saturated generalized gamma model was obtained, the Weibull model, the GG_λ-model and the spline model with μ=100 performed best, regarding a comparison of the estimated average subdistribution log-hazard ratios with the true value of -0.288 (MSE=0.038 for each model), the spline model with μ=1 and the saturated generalized gamma model were slightly worse (MSE=0.039 and MSE=0.042, respectively), while the spline model with the low value for the smoothing parameter (μ=0.01) lead to the highest variability in the estimated average subdistribution log-hazard ratios resulting in the highest MSE (MSE=0.066).

	Mean[a]	Median	$Q_{.05}$	$Q_{.95}$	Var (log)	MSE (log)
Weibull	0.750	0.749	0.549	1.022	0.038	0.038
GG_{λ}	0.748	0.743	0.547	1.017	0.039	0.039
$GG_{\lambda\tilde{\alpha}\nu}$	0.747	0.755	0.534	1.008	0.042	0.042
Spline (μ=0.01)	0.728	0.730	0.527	1.015	0.061	0.062
Spline (μ=1)	0.741	0.736	0.546	1.015	0.039	0.039
Spline (μ=100)	0.748	0.746	0.552	1.019	0.038	0.038

Table 7.6: Scenario IV - high censoring: Summary of estimated average subdistribution (log-)hazard ratios.
[a]Means of estimates for the average hazard ratio are exponentiated means of average log-hazard ratio estimates.

7.5 Summary and discussion of the simulation study

In this section a simulation study is described, that was performed in order to evaluate the ability of different mixture models to reflect given cause-specific and subdistribution hazard rates and hazard ratios. Parametric approaches assuming the conditional hazard rates of the mixture model to follow Weibull or generalized gamma distributions, either estimating one scale and shape parameter for both groups under investigation or allowing the shape and scale parameters to vary between both groups, that were presented in Section 4.4.2, and the approach estimating the conditional hazard rates using penalized B-spline basis functions, that was proposed in Section 5, were investigated and compared. The main findings regarding the estimated cause-specific or subdistribution hazard ratios were presented in this section, further results can be found in the Appendix (Section B.1). Competing risks data were simulated by application of algorithms presented in Sections 6.2 and 6.3 to generate scenarios with predefined cause-specific or subdistribution hazard rates and hazard ratios. Different true cause-specific or subdistribution hazard rates were used, leading to cause-specific hazard ratios either constant over time (Scenario I) or varying over time (Scenarios II to IV), and time-constant (Scenario IV) or time-dependent (Scenarios I to III) subdistribution hazard ratios. Different censoring distributions were considered in order to investigate the influence of the amount of censored observations on precision of the estimates and numerical properties of the estimation procedure. Results are illustrated by graphical display of summary statistics of the estimates obtained from the different models for prespecified timepoints. For scenarios with constant hazard ratios summary tables for estimated average (log-)hazard ratios are shown

As presented in the results section (Section 7.4), all methods performed pretty equally in scenarios with a true underlying cause-specific or subdistribution hazard ratio, which was constant over time and a low to moderate amount of censored observations. For the scenario with a time-constant subdistribution hazard ratio and a high amount of censored observations (about 75% of 1,000 observations) the spline model with a low value for the smoothing parameter (μ=0.01) showed a higher variability in the obtained estimates for the average log-subdistribution hazard ratio than the other models. When the true hazard ratios were time-dependent, the spline-based approach provided medians and exponentiated means of estimated log-hazard ratios, that were close to the true underlying hazard ratios, when the smoothing parameter was chosen adequately, while corresponding medians

and means obtained from the parametric approaches deviated from the true value, which was especially true for the Weibull and the generalized gamma mixture model, allowing only the location parameters to vary between groups.
For the generalized gamma models numerical maximization of the log-likelihood was not possible for a relevant number of datasets, as the maximization algorithm did not converge. This was especially noticeable for the generalized gamma approach, allowing all parameters to vary between the groups, in settings with a high amount of censoring, where mixture model regression coefficients, which are necessary for derivation of cause-specific and subdistribution hazard rates, could not be determined in up to almost one third of the investigated datasets. For the spline approach these numerical problems were only observed for the model with a low value of the smoothing parameter (μ=0.01), but the maximum number of datasets without adequately converged maximization algorithm was 5 (1.0%). For the spline models using higher values of the smoothing parameter, leading to smoother estimates of the conditional hazard rates, no numerical problems were observed. Numerical properties of the models under investigation for the considered scenarios are summarized in Table 7.7.
The simulation study revealed, that derivation of cause-specific and subdistribution hazard rates and consequently hazard ratios from a mixture model with conditional hazard rates estimated using the newly proposed spline approach, presented in Section 5, allows to reflect time-dependencies in the hazard rates and hazard ratios, when enough information is available. Penalizing the flexibility of the estimated conditional hazard rates, leading to smoother results, will improve the stability of the results, as models with a low value for the smoothing parameter showed a large variability, especially for later timepoints with only a low amount of information available. Parametric mixture models, which were proposed by Lau et al (2011) for flexible derivation of cause-specific and subdistribution hazard ratios from one mixture model, did not detect the true hazard ratios as adequately as the spline approaches, but estimates were less variable due to less parameters to be estimated, which was especially prevalent in simulation scenarios with a high amount of censored observations. For flexible parametric mixture models, as the generalized gamma mixture model allowing all three parameters of the conditional event time distributions to vary between groups, numerical problems were observed for the estimating procedure, especially for scenarios with time-dependent hazard rates and high amount of censored observations. So the newly proposed spline-based method appears to be an appealing approach for estimation of cause-specific and subdistribution hazard rates and hazard ratios from one model and for detection of time-dependencies in these quantities with better properties than the investigated parametric models in various scenarios.

		Parametric approaches			Spline approach		
	Cens.	Weib.	GG_λ	$GG_{\lambda\tilde{\alpha}\nu}$	μ=0.01	μ=1	μ=100
Scen. I	low	500 (100%)	500 (100%)	500 (100%)	500 (100%)	500 (100%)	500 (100%)
	med.	500 (100%)	498 (99.6%)	484 (96.8%)	500 (100%)	500 (100%)	500 (100%)
	high	500 (100%)	491 (98.2%)	419 (83.8%)	499 (99.8%)	500 (100%)	500 (100%)
Scen. II	low	500 (100%)	500 (100%)	492 (98.4%)	500 (100%)	500 (100%)	500 (100%)
	med.	500 (100%)	499 (99.8%)	465 (93.0%)	500 (100%)	500 (100%)	500 (100%)
	high	500 (100%)	484 (96.8%)	384 (76.8%)	498 (99.6%)	500 (100%)	500 (100%)
Scen. III	low	500 (100%)	500 (100%)	454 (90.8%)	500 (100%)	500 (100%)	500 (100%)
	med.	500 (100%)	500 (100%)	417 (83.4%)	500 (100%)	500 (100%)	500 (100%)
	high	500 (100%)	491 (98.2%)	342 (68.4%)	495 (99.0%)	500 (100%)	500 (100%)
Scen. IV	low	500 (100%)	500 (100%)	485 (97.0%)	500 (100%)	500 (100%)	500 (100%)
	med.	500 (100%)	498 (99.6%)	413 (82.6%)	500 (100%)	500 (100%)	500 (100%)
	high	500 (100%)	474 (94.8%)	341 (68.2%)	498 (99.6%)	500 (100%)	500 (100%)

Table 7.7: Numbers and proportions of converged algorithms for determination of maximum likelihood estimates for the investigated models in the different scenarios of the simulation study.

Chapter 8

Application to data from a clinical cohort study

The competing risks regression models described in Section 4 and the approach for estimation of cause-specific and subdistribution hazards from a mixture model, using P-splines to model the conditional hazard rates, presented in Section 5, were applied to a dataset from a clinical cohort study. Background information on the data is given in Section 8.1.1 and summaries of event time data are presented in Section 8.1.2. Details on the applied competing risks regression models and results obtained from the analyses are shown in Section 8.2. In Section 8.3 the hazard rates and ratios smoothly estimated from a mixture model are shown and details on the estimation procedure are given. Application of competing risks regression models including details on the applied methods and results obtained from analyses, which are presented in Section 8.2, were published in Haller et al (2013). Application of the mixture model using the P-spline approach and according results shown in Section 8.3 are described in a manuscript, which was under review when this work was finalized.

8.1 Description of the data

8.1.1 Study description

The presented methods were all applied to a dataset collected in a cohort study, which was conducted in the Klinikum rechts der Isar and in the German Heart Centre Munich, both located in Munich, Germany, between January 1995 and March 2005. A total of 2343 patients, who survived an acute myocardial infarction (MI) at an age of 75 years or younger, were included in the study. The analysed data are presented in Bauer et al (2006, 2009) and Barthel et al (2003), including medical details and a more substantial description of the study cohort. Two of the patients were excluded from the analyses due to missing values, so the results presented are based on the evaluation of 2341 individuals. Patients were planned to be followed for five years. Time from myocardial infarction to death and type of death (cardiac or non-cardiac reason) were documented. Patients were stratified regarding their risk for cardiac death. Patients with a left ventricular ejection fraction (LVEF) of less than 30% and patients with an LVEF of more than 30%, but severe autonomic failure (SAF), were specified to be of high risk for cardiac death (n=236), patients with an LVEF of more

than 30% and no SAF to be of low risk (n=2105).
1140 patients were followed for five years, so the median follow-up time was five years assessed by the inverse Kaplan-Meier method (Schemper and Smith, 1996). About 75% of the patients were followed for at least three years. Patients lost to follow-up or retreating from the trial were considered as censored observations. During follow-up 181 of the 2341 patients died, 104 of them from cardiac reasons (55 sudden cardiac deaths), 77 patients died from other causes or types of death were not specified (n=14). For ease of analysis and interpretation these 77 patients were defined to have died from non-cardiac reasons.

8.1.2 Summary of observed events

The cumulative incidence functions for cardiac death and non-cardiac death for the whole study population and for both risk groups were estimated as shown in Equation 3.17, confidence intervals were derived using the estimator for the asymptotic variance proposed by Aalen (1978a). Estimates were calculated using the function *cuminc* of the R package *cmprsk* (Gray, 2010).
The estimated probability of dying in the first five years after myocardial infarction for the whole study population was 9.2% (95% confidence interval 7.9% to 10.5%) with a cumulative incidence for cardiac death of 5.1% (95% ci 4.2% to 6.1%) and for non-cardiac death of 4.1% (95% ci 3.1% to 5.0%). For cardiac death a large difference between the cumulative incidence functions of both groups was observed, with an estimated five year probability for cardiac death of 27.5% (95% ci 21.1% to 33.9%) in the high risk group and 2.7% (95% ci 1.9% to 3.4%) in the low risk group. Estimated five year probabilities for non-cardiac death also differed between both risk groups, but the observed difference was smaller, with 10.9% (95% ci 6.3% to 15.7%) for the high risk group and 3.3% (95% ci 2.4% to 4.2%) for the low risk group.
Estimates of the cumulative incidence functions five years after myocardial infarction with 95% confidence intervals for death from any kind and for both types of death are presented in Table 8.1 for the whole study population and stratified for risk groups. In Figure 8.1 non-parametric estimates of the cumulative incidence functions for the two competing types of event are presented. The sum of the cumulative incidence functions for both types of event, presented as solid line in Figure 8.1 (a), can be interpreted as an estimate for one minus overall survival (see Equation 3.4).

	Cardiac death		**Non-cardiac death**	
	$\underline{\hat{F}}_{card.}(5\,years)$	95% ci	$\underline{\hat{F}}_{non\text{-}card.}(5\,years)$	95% ci
Overall	5.1%	4.2% to 6.1%	4.1%	3.1% to 5.0%
Low risk	2.7%	1.9% to 3.4%	3.3%	2.4% to 4.2%
High risk	27.5%	21.1% to 33.9%	10.9%	6.3% to 15.7%

Table 8.1: Estimated cumulative incidences five years after myocardial infarction with 95% confidence intervals.

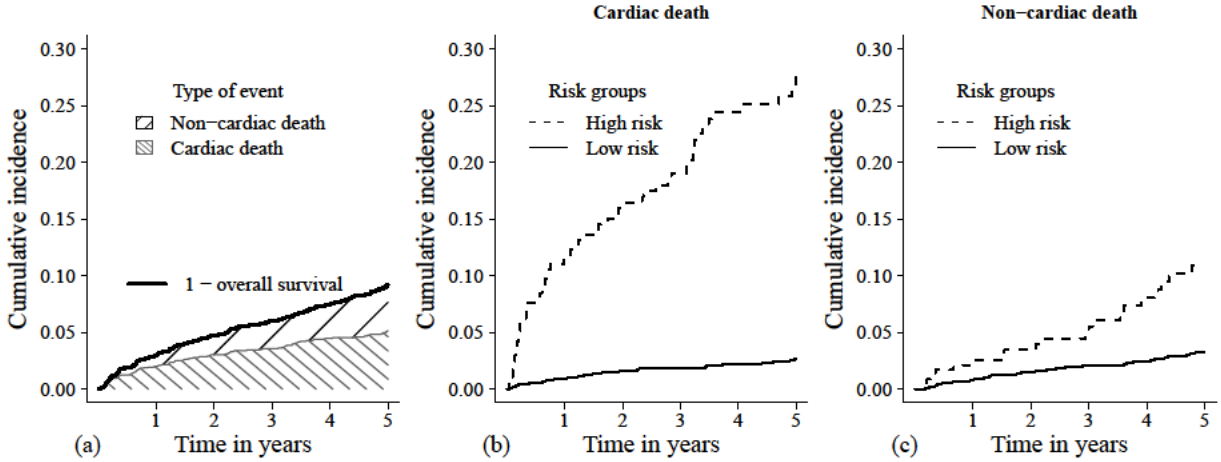

Figure 8.1: Estimated cumulative incidence functions for cardiac and non-cardiac death for the whole study population. Cumulative incidences for both types of event sum up to one minus overall survival (a). Comparison of high and low risk group regarding incidences of cardiac (b) and non-cardiac death (c).

8.2 Application of regression models

In order to evaluate the risk stratification, the groups of patients defined as being of high risk for cardiac death and of low risk were compared using the regression models presented in Section 4, adjusting for age, considered as a binary covariate indicating whether a patient's age was 65 years or higher, and diabetes status.

8.2.1 Cause-specific hazards regression

The effect of risk group allocation on the cause-specific hazards adjusted for age and diabetes was analysed. As an investigation of Schoenfeld residuals provided no evidence against the assumption of proportionality for both types of event (not shown), a Cox regression model was fit for each type of failure to estimate the effect of the three covariates on the cause-specific hazards, in order to described the whole competing risks process. The R function *coxph* from the library *survival* (Therneau, 2011) was used for estimation of the regression coefficients, treating patients that failed from a competing event as censored observations as described and discussed in Section 4.1. Risk group, diabetes, and age had a significant effect on the cause-specific hazards for both types of event (results are shown in Table 8.2). A cause-specific hazard ratio between the high risk and the low risk group for cardiac death (HR_c^{cs}) of 10.53 (95% ci 7.10 to 15.64) was estimated. This indicates an about ten times higher risk of dying from a cardiac event for patients with an LVEF $\leq 30\%$ or with SAF compared to patients with an LVEF $> 30\%$ and no SAF. The analysis for non-cardiac deaths revealed an increased risk for patients from the high risk group, too, but the effect was much smaller ($HR_{nc}^{cs} = 2.89$, 95% ci 1.73 to 4.85). Age had a greater influence on the cause-specific hazard for non-cardiac death with a cause-specific hazard ratio of 3.69 (95% ci 2.29 to 5.91) compared to 1.60 (95% ci 1.09 to 2.40) for cardiac events. The cause-specific hazard for a patient with diabetes was about twice as high as for a patient without diabetes for both types of event.

Cardiac death				
	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. error	p-value
Risk group	2.36	10.53	0.20	<0.001
Diabetes	0.72	2.06	0.21	0.001
Age $\geq$ 65	0.48	1.60	0.20	0.016

Non-cardiac death				
	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. error	p-value
Risk Group	1.06	2.89	0.26	<0.001
Diabetes	0.70	2.01	0.25	0.005
Age $\geq$ 65	1.28	3.69	0.24	<0.001

Table 8.2: Results of the cause-specific hazards regression models.

Cumulative incidence functions were predicted from the Cox regression models following Equation 4.7 for both risk groups, using the mean of diabetes, i.e. the proportion of patients with diabetes (17.6%), and the mean of the indicator variable for age, i.e. the proportion of patients being at least 65 years of age (30.2%). Cause-specific baseline hazards, which are required for calculation of cumulative incidence functions, were derived using the generalized Breslow estimator shown in Equation 4.6. The predicted cumulative incidence curves are displayed in Figure 8.4 (a).

8.2.2 Subdistribution hazards regression

A proportional subdistribution hazards model as described in Equation 4.8 was fit to the data in order to assess the influence of risk group, diabetes, and age on the subdistribution hazards for both types of event. The analysis was performed using the function *crr* in the R library *cmprsk*. Due to conceptual problems present when proportional subdistribution hazard models are fit for both types of event, as shown by Beyersmann et al (2012), results from the subdistribution hazards models have to be interpreted as time-averaged effects (Grambauer et al, 2010).
Results of the two regression models investigating the influence of the covariates on the subdistribution hazards are shown in Table 8.3 for both types of failure. Effects on the subdistribution hazards can be translated directly to effects on the cumulative incidence functions. A comparison between the high risk and the low risk group revealed an average subdistribution hazard ratio for cardiac death (HR_c^{sd}) of 10.21 (95% ci 6.91 to 15.08), indicating a much higher incidence of cardiac events for patients categorized to be of high risk compared to low risk patients. The effect of the risk group allocation was weaker for non-cardiac death ($HR_{nc}^{sd} = 2.31$, 95% ci 1.39 to 3.97). Effects of diabetes were similar for both types of failure with a higher subdistribution hazard for patients suffering from diabetes, whereas age had a higher effect on the subdistribution hazard of non-cardiac death. Cumulative incidence functions for cardiac death comparing high and low risk group at mean of age and diabetes are shown in Figure 8.4 (b), which is displayed on page 112.

	Cardiac death			
	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. error	p-value
Risk group	2.32	10.21	0.20	<0.001
Diabetes	0.68	1.98	0.21	0.001
Age $\geq$ 65	0.47	1.60	0.20	0.017

	Non-cardiac death			
	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. error	p-value
Risk Group	0.84	2.31	0.28	0.002
Diabetes	0.62	1.85	0.24	0.011
Age $\geq$ 65	1.28	3.58	0.25	<0.001

Table 8.3: Results of the subdistribution hazards (Fine and Gray) regression models.

Due to the high amount of censored observations and the low number of competing events, leading to similar risk sets for estimation of cause-specific and subdistribution hazards (see Equations 3.6 and 3.11, and Figures 3.2 and 3.3), both hazard based regression models lead to similar results for the data example.

8.2.3 Semi-parametric mixture model assuming proportional conditional hazard rates

For analysis of the data using a semi-parametric mixture model, the approach proposed by Ng and McLachlan (2003), which is described in Section 4.4.3, was applied, so no assumption for the failure time distribution for a given type of event had to be made, but conditional hazard rates were assumed to be proportional. Parameter estimates were obtained via an expectation-conditional maximization (ECM) algorithm, where parameters are estimated iteratively by altering the expectations of failure types for censored observations and consequently the expectation of the log-likelihood function, given the observed data and the current parameter estimates (E-step), and maximization of the log-likelihood given the observed data and the expected failure-type probabilities for censored observations (M-step). These steps were repeated until the sum of the absolute differences for all regression coefficients between two consecutive steps was smaller than 10^{-5}. Different starting values were used to avoid finding a local maximum, but all computations led to the same results.
Five hundred bootstrap samples were generated to estimate confidence intervals for the regression coefficients. As described by Ng and McLachlan, subsamples were randomly drawn with replacement from patients experiencing cardiac death, from patients failing from non-cardiac death, and from censored individuals according to the numbers observed in the original dataset. Results of the analysis are presented in Table 8.4.
The coefficients of the logistic regression model, assessing the marginal event type distribution, indicate that high risk patients were more likely to die from cardiac events ($OR = \exp(2.22) = 9.21$, 95% bootstrap ci 0.24 to 52.98). For a low risk patient aged at

Event types		
	Cardiac	
	$\hat{\beta}$	95% ci (bootstrap)
Constant	-2.30	-3.92 to 1.65
Risk group	2.22	-1.44 to 3.97
Diabetes	-0.43	-2.00 to 1.82
Age $\geq$ 65	0.96	-1.45 to 2.66

Event times				
	Cardiac		**Non-cardiac**	
	$\hat{\beta}$	95% ci (bootstrap)	$\hat{\beta}$	95% ci (bootstrap)
Risk group	0.88	-0.92 to 3.19	1.76	-0.21 to 2.76
Diabetes	1.17	-0.44 to 2.02	0.52	-0.49 to 2.07
Age $\geq$ 65	-0.25	-1.60 to 1.02	1.54	0.71 to 2.54

Table 8.4: Regression coefficients obtained from the mixture model analysis with 95% confidence intervals based on 500 bootstrap samples. Regression coefficients for the marginal event type distribution model are presented in the upper table, coefficients for the conditional event time models in the lower table.

least 65 years and having no diabetes a probability of dying from a cardiac event of 20.7% was estimated, for a person of the same age, who is also free of diabetes, but who was identified to be of high risk, the predicted probability increases to 70.7%. For both types of failure, patients from the high risk group tended to survive for a shorter time period, as their estimated risk for failing from the given type of event is increased (hazard ratios of exp(0.88)=2.41 and exp(1.76)=5.81). Due to the low number of events and the large amount of censoring, confidence intervals derived from the bootstrap samples are very wide.

8.2.4 Vertical Modelling

In the vertical modelling approach, which was proposed by Nicolaie et al (2010) and is described in Section 4.6, patterns for the occurrence of events in the course of time can be investigated. Marginal survivor functions for both risk groups, adjusted for age and diabetes, were estimated from a Cox regression model using the R function *coxph* of the *survival* library, considering time to death irrespective of the event type, and are presented using the mean of diabetes and the mean of age derived from the whole study population (Figure 8.2 a).
In order to estimate relative hazards of the event types in the course of time, a logistic regression model was fit considering all uncensored subjects (n=181). Time, risk group, diabetes, and age were included as covariates, an indicator variable giving one, if the observed event was death from a cardiac reason and zero for death from a non-cardiac

	$\hat{\beta}$	Std. error	p-value
Constant	0.75	0.61	0.218
Risk group	1.27	0.91	0.165
Diabetes	-0.08	0.36	0.825
Age $\geq$ 65	-0.65	0.34	0.053
bs1(time)	-0.76	1.96	0.698
bs2(time)	-1.37	1.69	0.419
bs3(time)	-0.69	1.14	0.545
bs1(time)$\times$risk group	1.08	3.25	0.740
bs2(time)$\times$risk group	-0.64	2.56	0.802
bs3(time)$\times$risk group	-0.46	1.82	0.800

Table 8.5: Results from the vertical modelling approach. "bs" denotes B-Spline components of time, the last three lines show the interaction terms between the B-Spline components and risk group.

reason, was used as dependent variable. Cubic B-spline functions were used to estimate the effect of time smoothly. As proposed by Nicolaie et al (2010), interaction terms between risk group and the smooth functions of time were considered to allow for different patterns in both groups.
Calculations were conducted using the function *glm* with flexible B-splines, which are incorporated in the *splines* library. Coefficients obtained from the logistic regression model, estimating the probability of occurrence of a cardiac event given any event was observed, are presented in Table 8.5. As interpretation of regression coefficients is difficult due to the use of B-spline functions and interaction terms, estimated relative hazards are displayed graphically for both types of event for the high risk group in Figure 8.2 (b) and for the low risk group (c).

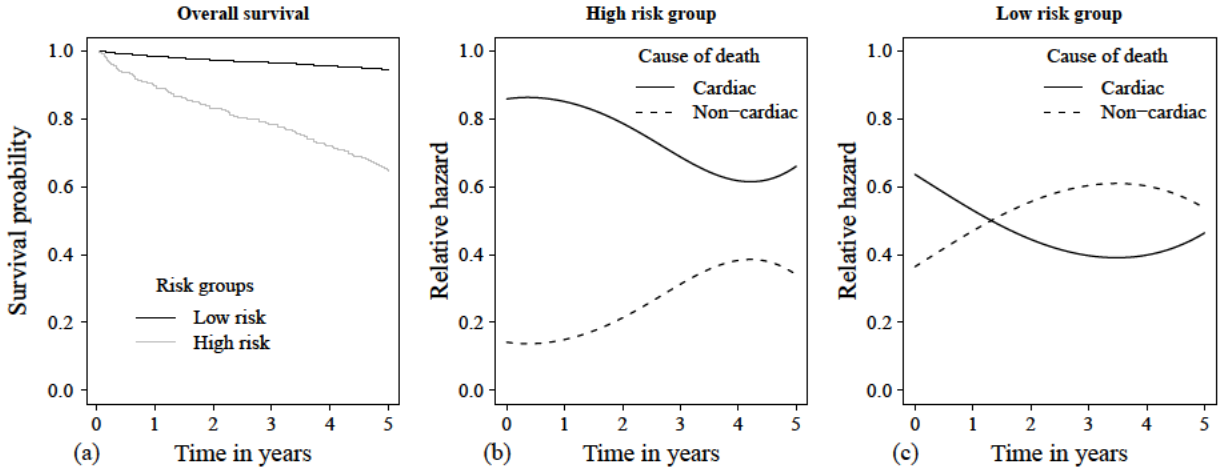

Figure 8.2: Results of the vertical modelling approach: Survivor functions adjusted for age and diabetes (a), relative hazards for the high risk group (b) and relative hazards for the low risk group (c).

The estimated probability for death from any type, adjusted for age and diabetes, is higher in the high risk group compared to the low risk group (Figure 8.2 a). For a high risk patient the probability of dying from a cardiac event, given the patient dies at a certain time t, is substantially higher than the probability for a non-cardiac event for all timepoints (b), whereas both types of event seem to appear with a similar probability in the low risk group (c). For both types of event the probability for a cardiac event, given any event occurred, seems to decrease slightly over time, indicating a higher relative hazard for cardiac events in the first one to two years after myocardial infarction.

8.2.5 Pseudo-observation approach

In order to analyse the data using the approach proposed by Klein and Andersen (2005), which is sketched in Section 4.7, pseudo-values were estimated for each individual, which were used as response in a GEE model. In a first step, the cumulative incidence function for cardiac death was estimated for 21 different points in time (three month intervals equally spaced from baseline to five years of follow-up) for the whole dataset as described in Equation 3.17. The procedure was repeated for all prespecified timepoints leaving out each subject once. Pseudo-observations were calculated from these 2341×21 estimates following Equation 4.41. Examples for pseudo-observations obtained from the observed data are displayed in Figure 8.3. The patient displayed in the left picture (a) died from a cardiac reason after 2.57 years, the patient in Figure 8.3 (b) was censored after 3.79 years, and the patient in the right picture (c) died from a non-cardiac reason after 2.07 years. For better illustration a different scale was used for Figures (b) and (c) compared to Figure (a). Due to the large amount of patients followed for five years without any critical event, the pseudo-value approach does not affect the weights of censored individuals heavily, but patients experiencing an event of interest, as the patient illustrated in Figure 8.3 (a), will obtain pseudo-values larger than one for time points after their time of cardiac death, the value depending on the event time.

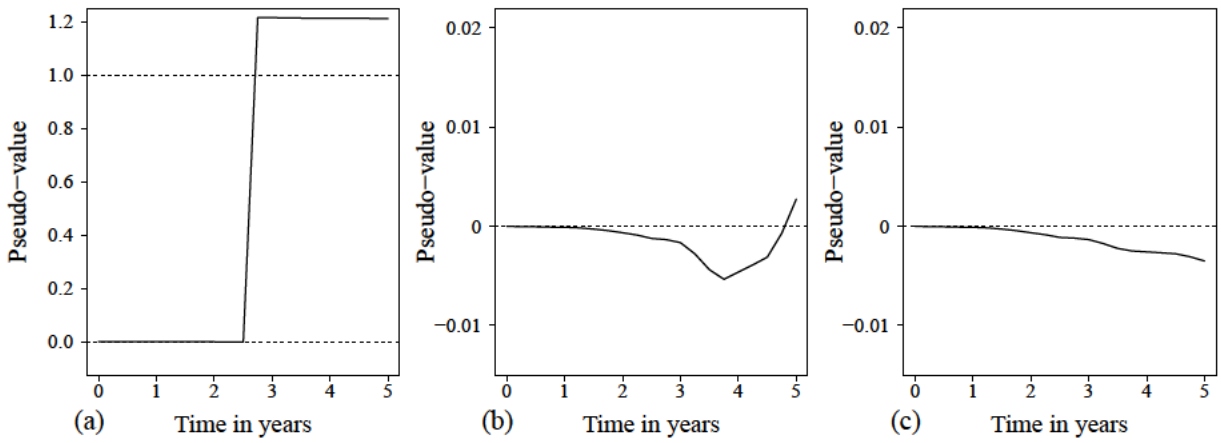

Figure 8.3: Examples for derived pseudo-values: Cardiac death after 2.57 years (a), censored after 3.79 years (b), non-cardiac event at 2.07 years (c). Different scales are used for (b) and (c) compared to (a).

	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. error	p-value
Constant	-6.81	0.00	0.35	<0.001
Risk group	2.36	10.59	0.22	<0.001
Diabetes	0.81	2.25	0.25	0.001
Age $\geq$ 65	0.53	1.70	0.26	0.043
Time = 3 months	1.04	2.83	0.30	<0.001
Time = 6 months	1.21	3.35	0.26	<0.001
Time = 9 months	1.65	5.21	0.23	<0.001
Time = 12 months	1.73	5.64	0.22	<0.001
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Time = 57 months	2.68	14.59	0.18	<0.001
Time = 60 months	2.74	15.49	0.18	<0.001

Table 8.6: Regression coefficients obtained by the pseudo-value approach. Coefficients are not shown for all time points. Skipped coefficients, which are needed for estimation of the cumulative incidence functions, are monotonously increasing. $\exp(\hat{\beta})$ for risk group, diabetes, and age can be interpreted as subdistribution hazard ratio.

These pseudo-values were used as dependent variable in a GEE model, to account for multiple observations of the same subjects. Age, diabetes, and 20 dummy variables indicating the timepoint were included as covariates. The independence working covariance matrix was used in the GEE model, which was suggested by Klein and Andersen (2005). The influence of the covariates of interest on the pseudo-values was estimated using a complementary log-log (cloglog) link between the response and the linear predictor, applying the function *geese* of the R library *geepack* (Højsgaard et al, 2005), so the estimated coefficients can be interpreted as logarithms of subdistribution hazard ratios. The results of the GEE model are presented in Table 8.6.

Effects observed in the pseudo-value approach are similar to those obtained in the Fine and Gray model and can be interpreted analogously as an effect on the subdistribution hazard, translating to an effect on the cumulative incidence function (Klein and Andersen, 2005). A subdistribution hazard ratio comparing the high risk to the low risk group of $\exp(2.36) = 10.59$ was estimated (95% ci 6.88 to 16.28). As described by Andersen and Perme (2010), the standard errors obtained in the pseudo-value approach are higher than those in the Fine and Gray regression model.

Regression coefficients for the different timepoints specified for calculation of the pseudo-observations, which are partly presented in Table 8.6, are not of major interest, but are necessary for estimation of the cumulative incidence function. The estimated cumulative incidence function, derived from results of the pseudo-observation approach, which is shown in Figure 8.4 (c), is similar to the cumulative incidence functions obtained from the cause-specific hazards regression or the subdistribution hazards regression. Steps of the function are obtained for each of the timepoints specified for the estimation of pseudo-values (three month intervals between t=0 and t=5 years).

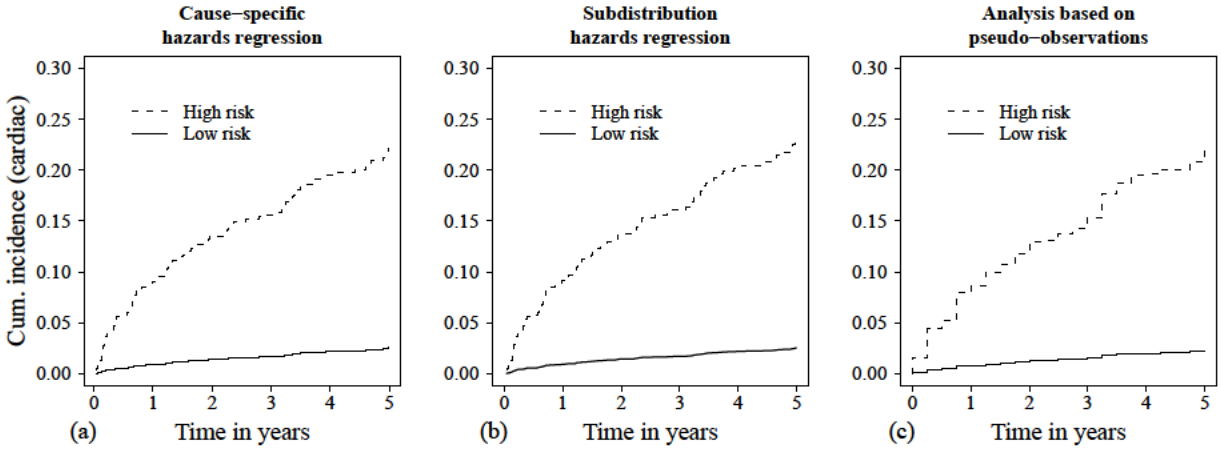

Figure 8.4: Estimated cumulative incidence functions for cardiac death using the cause-specific hazards regression (a), the subdistribution hazards regression (b), and the analysis based on pseudo-observations (c). For the cause-specific and the subdistribution hazards regression steps are obtained for each observed event time with an event of interest, for the pseudo-observations approach for each timepoint specified for calculation of pseudo-values.

8.3 Application of the newly proposed spline-based mixture model approach

The spline-based approach presented in Section 5 was applied to the data, using different values for the smoothing parameter (μ=1, μ=100). In accordance to the method description in Section 5, comparison of risk groups was conducted without consideration of further covariates. A number of five interior knots was used for determination of the basis functions and knots were spaced depending on observed event time quantiles as described for analysis of the simulated data (Section 7.2). Coefficients of the mixture model were estimated by numerical maximization of the log-likelihood using the function *nlm*. Cause-specific and subdistribution hazards and hazard ratios were derived from the mixture models as presented in Section 4.5. For both spline models pointwise 90% bootstrap confidence intervals were estimated. Bootstrap samples were drawn from the subsamples of individuals failing from a cardiac event, failing from a non-cardiac event, and from censored observations, with the numbers of samples equalling the numbers observed in the original data as described by Ng and McLachlan (2003). Five hundred bootstrap samples were drawn and confidence intervals were estimated for a set of prespecified timepoints as 5% and 95% quantiles of the estimated cause-specific hazard rates and ratios derived from the bootstrap samples.
Estimates of the cause-specific hazards and hazard ratios with 90% confidence intervals obtained from the spline-based approach for smoothing parameters of μ=1 and μ=100 are presented in Figure 8.5. In a Cox-type regression model for the cause-specific hazards (Prentice et al, 1978, described in Section 4.1), considering risk group, but no other covariates, and assuming the cause-specific hazard ratio to be constant over time, a cause-specific hazard ratio for the event of interest of 12.3 was estimated (90% confidence interval: 8.9 to 17.0). The hazard ratio derived from the Cox-type regression model is displayed in the according pictures of Figure 8.5 (c and d) as dashed line.
Estimates for the subdistribution hazard rates and hazard ratios with 90% confidence inter-

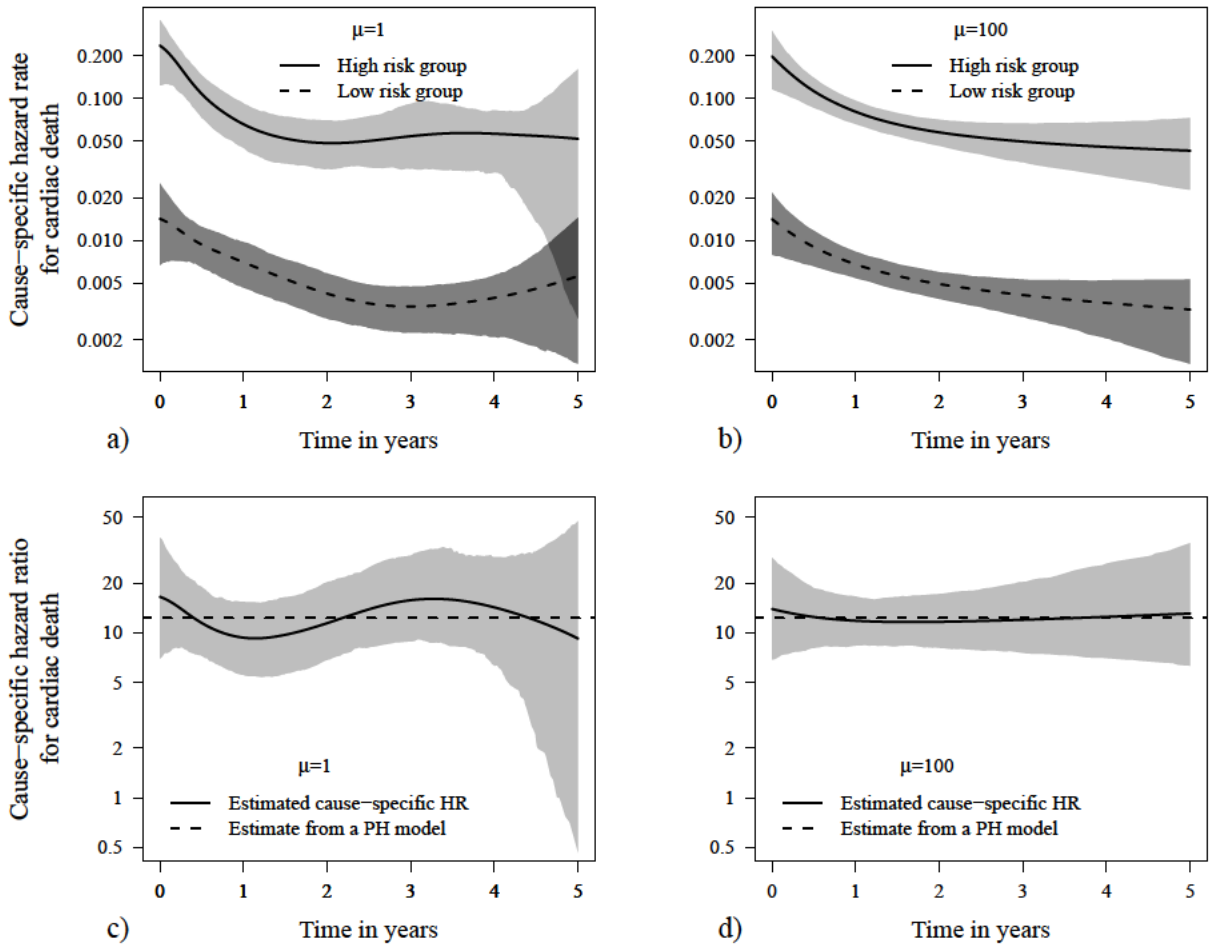

Figure 8.5: Cause-specific hazard rates (top row) and cause-specific hazard ratios (bottom row) derived from a mixture model using the spline approach for estimation of the conditional hazard rates with smoothing parameters of μ=1 (left column) and μ=100 (right column). Grey areas illustrate 90% confidence intervals derived from 500 bootstrap samples.

vals are displayed in Figure 8.6. A Fine and Gray regression model, assuming proportional subdistribution hazards between both risk groups, was fit to the data (Fine and Gray, 1999, described in Section 4.2), providing a subdistribution hazard ratio of 11.9 (90% ci 8.7 to 16.5). The estimated subdistribution hazard ratio is shown in Figure 8.6 (c and d) as dashed line, in order to compare the result of the spline approach to that obtained from a proportional subdistribution hazards model.

As revealed in previous analyses, individuals defined as being of high risk for cardiac death were found to have higher cause-specific and subdistribution hazards compared to low risk patients. In this data example estimated cause-specific and subdistribution hazard rates and consequently hazard ratios are similar, due to the small number of competing events, resulting in similar values for $S(t|X)$ and $1-\underline{F}_{card.}(t|X)$ (see Equations 4.30 and 4.32). For both types of hazard, there is no evidence for invalidity of the proportionality assumption, and results of the spline approach are similar to those obtained from the proportional hazards models. In comparison to the proportional hazards models, additional information

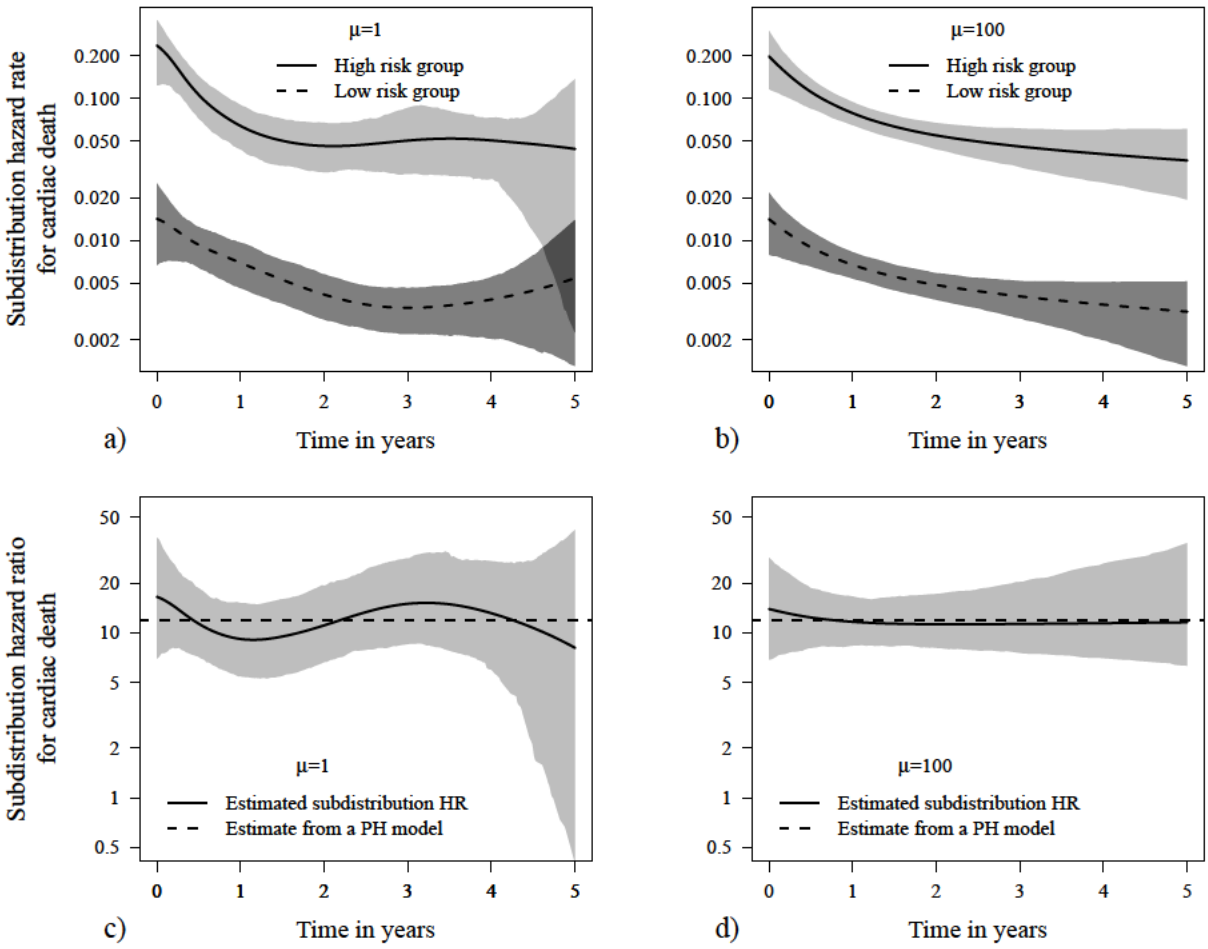

Figure 8.6: Subdistribution hazard rates (top row) and subdistribution hazard ratios (bottom row) derived from a mixture model using the spline approach for estimation of the conditional hazard rates with smoothing parameters of μ=1 (left column) and μ=100 (right column). Grey areas illustrate 90% confidence intervals derived from 500 bootstrap samples.

on the cause-specific and subdistribution hazard rates is obtained. The estimated hazard rates indicate an increased risk for cardiac death in the first two years after the myocardial infarction for both risk groups. It has to be considered that confidence intervals might not be valid when penalized spline methods are used and might be falsely narrow for a high value of the smoothing parameter.

8.4 Summary of results and applicability

In this section the methods for analysis of competing risks data presented in Section 4 and Section 5 were applied to data from a clinical cohort study investigating risk stratification for cardiac death in patients that survived a myocardial infarction at an age of 75 years or younger. The risk stratification appears to do well, irrespective of the model used for analysis. When comparing the results obtained from the different models, it has to be considered, that the models use different quantities and underlie different assumptions.

The most popular models, the cause-specific hazards regression considering a Cox-type regression model and the Fine and Gray regression model, being a Cox-type regression model for the subdistribution hazard, revealed very similar results in this data example with a cause-specific hazard ratio of 10.53 between the two risk groups after adjustment for diabetes and age and a corresponding subdistribution hazard ratio of 10.21, due to the high amount of censored observations and the low number of competing events. This will generally not be the case, as discussed in Section 4.3. In order to estimate the cumulative incidence functions for cardiac death, using results obtained from the cause-specific hazards regression approach, models for cardiac and non-cardiac type of death have to be estimated separately and regression coefficients for both types of event have to be considered. For the subdistribution hazards regression model the cumulative incidence functions are monotonously linked to the coefficients obtained from a regression model for the event of interest (cardiac death) and estimation does not rely on a model for the competing event, as shown in Equation 4.12.
The semi-parametric mixture model approach proposed by Ng and McLachlan (2003), which is discussed in Section 4.4.3 and which was applied to the dataset (results in Section 8.2.3), uses a factorization of the joint distribution of event times and event types by modelling the marginal event type distribution and the conditional event time distributions, given the type of event. Therefore, estimates of event probabilities can be obtained from the mixture model approach under the assumption, that models for the conditional event time distributions also hold true for timepoints after the end of the study. Applying the semi-parametric mixture model to the study data revealed an odds ratio of 9.21, indicating a higher probability for a cardiac event for patients identified as being of high risk compared to low risk patients, but the 95% bootstrap interval was very wide (0.24 to 52.98) due to the high amount of censored observations.
The vertical modelling approach proposed by Nicolaie et al (2010) uses another factorization of the joint distribution for event times and event types by considering the marginal event time and the conditional event type distribution. Graphical display of the relative hazards, representing the estimated conditional probabilities for the different event types given any event was observed, was recommended. Estimates of the relative hazards can be derived from a logistic regression model using time and other measures of interest as independent variables. Incorporation of spline functions and interaction terms allows flexible estimation of the relative hazards, if enough information is available. For the investigated example the influence of time on the event type probabilities was assessed using B-spline functions and an interaction term between time and risk group was used. Graphical display of the relative hazards over time for both risk groups revealed a higher risk for cardiac death than for non-cardiac death in the high risk group and similar risks for both types of event for the low risk group. For both groups the relative hazard for cardiac death appeared to decrease over time. The marginal event time distribution was also illustrated and should always be considered, in order to assess the overall risk for any event and to avoid overinterpretation of relative hazards for time intervals with a low number of observed events. In this example the probability of failing from any event five years after myocardial infarction is substantially higher for a high risk patient compared to a low risk patient.
The pseudo-value approach, for which the theoretical background is presented in Section 4.7, gave results similar to the subdistribution hazards regression, with a subdistribution

hazard ratio of 10.59. This result was to be expected, as the complementary log-log link, which was used in the GEE model, leads to a link between linear predictor and cumulative incidence function as in the proportional subdistribution hazards regression model. Andersen and Perme (2010) discussed, that application of the pseudo-value approach with a complementary log-log link does not have any advantages over the subdistribution hazards regression approach proposed by Fine and Gray, as it leads to similar results with larger standard errors, which could also be observed for this data example, but that the pseudo-value approach in general allows more flexible modelling in settings where the proportional hazards assumption does not hold.
The hazard based approaches, the cause-specific and the subdistribution hazards regression, could be applied most easily using standard functions available e.g. in the statistical software packages R or SAS. Application of the semi-parametric mixture model approach required implementation of the ECM algorithm presented by Ng and McLachlan (2003), as no software routines were available. A sketch of the R code used can be found in the Appendix (Section C.3). The vertical modelling approach could be applied using a standard logistic regression model for estimation of the relative hazards including splines and interactions, and an estimate for the marginal event time distribution could be obtained using common survival methods. Little implementation was necessary to derive the relative hazards from the regression coefficients of the logistic regression model and for adequate graphical presentation. For the pseudo-value approach R and SAS functions were provided by Klein et al (2008). These functions can be used to generate the pseudo values, which are considered as response in a GEE model, that can be applied using available functions in R or SAS.

Chapter 9

Discussion and conclusion

9.1 Discussion

In this work the analysis of event time data in the presence of competing risks, i.e. when subjects can fail from one out of two or more mutually exclusive types of event, is described and discussed. Adequate analysis of competing risks data is relevant for various applications. In medical research time to a certain cause of death might be of major interest in order to assess efficacy of a therapy or the predictive or prognostic effect of a certain risk factor, with other causes of death being competing risks. In a recent article by Koller et al (2012) the availability of methods for adequate analysis of competing risks data, assessed by reviewing the most important biostatistical journals, and the application of competing risks methods for analysis and presentation of clinical data in the highest ranked medical journals was investigated. While a large number of methods for analysis of competing risks data was found in the statistical literature, adequate application and description of these methods was only found in a relatively small proportion of the investigated articles in medical journals. Nevertheless, problems and pitfalls present in the competing risks setting have become more recognized in the medical community in recent years, which is indicated by publication of articles in medical journals, describing and discussing adequate analysis of competing risks data (e.g. Dignam and Kocherginsky, 2008; Berry et al, 2010; Roobol and Heinsdijk, 2011).
The description of competing risks data, especially in medical research, is often conducted using standard survival methods, which were developed for the analysis of time-to-event data with one possible endpoint, without adequate consideration of competing risks. Results of the "naïve" Kaplan-Meier estimator, treating competing events as censored observations, are presented in many articles as estimates for the probability of being free of the event of interest up to a given time. As this violates a fundamental assumption of the Kaplan-Meier estimator, namely the independence between event times and censoring times, the event probabilities are overestimated by that procedure, which was discussed in various articles (e.g. Putter et al, 2007; Andersen and Keiding, 2012) and is illustrated for a data example in this work. Unbiased estimates for the probability of failing from a certain event up to a given time, which is represented by the cumulative incidence function in the competing risks setting, can be obtained by application of the Aalen-Johansen estimator, that relies on the so called cause-specific hazard rates. The cause-specific hazard rates are the natural adaptations of the common hazard rate for the competing risks setting and

many authors argue for the use of the cause-specific hazard rates, as these "completely determine the competing risks process" (Beyersmann et al, 2009) and allow "for a 'direct' formulation of the effect of exposure on the instantaneous forces that drive the patients remaining at risk at each time point t, that is, those without any prior event" (Koller et al, 2012). Since the probability for an event of interest is not directly related to the cause-specific hazard for the event of interest, but relies on the cause-specific hazards for all types of event, as risks on competing events have an influence on the number of patients at risk, an alternative hazard rate was constructed by Gray (1988), the so called subdistribution hazard rate. For the subdistribution hazard rate the risk set is adapted by keeping individuals in the risk set, that failed from an event other than the event of interest. This results in a direct relationship between the subdistribution hazard rate and the cumulative incidence function as known from standard survival analysis. For estimation of the subdistribution hazard rate in the presence of censored observations, a potential censoring time has to be determined for each individual to obtain unbiased estimates. While use of the subdistribution hazard rate appears appealing due to its direct relationship to the cumulative incidence function, its use was argued against, because of the unintuitive risk set formulation (Andersen and Keiding, 2012).

Different methods for analysis of event time data with multiple types of event were proposed, in order to estimate the joint distribution of event times to different types of event, assuming a correlation structure for times to different event types, as e.g. models for exponentially distributed event times with two possible types of event, which were introduced and discussed by Cox (1959). As in a classical competing risks setting event types are mutually exclusive, the correlation structure cannot be estimated from observable data and the assumption for the correlation structure cannot be verified. Various models for estimation of possible ranges for the correlation structure based on the marginal event time distributions were developed (see e.g. Peterson, 1976; Klein and Moeschberger, 1988), but these were not presented in this work, as only models, that are estimable from observable data and that do not rely on unverifiable assumption, are considered.

Various regression models for competing risks data were introduced since the end of the 1970s, that can be estimated from observable data. While Prentice et al (1978) proposed to analyse competing risks data by modelling the cause-specific hazard rates, Fine and Gray (1999) introduced a regression model for the subdistribution hazards. The two hazard based models are the most popular regression models for competing risks data. Alternative approaches for regression analysis were proposed by Larson and Dinse (1985) and Nicolaie et al (2010), who proposed to split the joint distribution of event types and event times in one marginal and one conditional distribution. Larson and Dinse (1985) represented the joint distribution by the product of the marginal event type distribution and the conditional event time distributions, given the type of event, which is known as mixture model approach in the competing risks setting. Nicolaie et al (2010) introduced the so called vertical modelling approach, assessing the marginal event time distribution and the conditional event type distribution given the time of event, which provides relative hazards for the different event types over time. Andersen et al (2003) introduced a calculation technique for estimation of covariate effects on event probabilities in multi-state models using pseudo-values, that are derived by jackknife estimates from the original data. This approach was later described for the competing risks setting (Klein and Andersen, 2005), which is a special case of a multi-state model.

The different regression models are presented in this work and were applied to a dataset from a clinical cohort study, investigating a risk stratification for cardiac death in patients that survived a myocardial infarction. Due to the high amount of censored observations, results from the two hazard-based methods were similar in that example, which does not have to be the case, as the effect on different quantities is investigated. The mixture model approach, which focusses on the marginal event type distribution, revealed a substantially higher probability of dying from a cardiac event for patients, defined as being of high risk for cardiac mortality, than for low risk patients. In the vertical modelling approach a higher relative hazard for death from a cardiac reason than from a non-cardiac reason was estimated for the high risk group, while relative hazards for both types of death were similar in the low risk group. For both groups the relative hazard for death from a cardiac reason was found to decline over time. The pseudo-value approach with a complementary loglog-link gave results similar to the hazard-based regression models. A more comprehensive discussion on the results of the data analysis and applicability of the different regression approaches can be found in Section 8.4 of this work and in a publication, in which the competing risks regression models are described and compared, and results from data application are shown (Haller et al, 2013).
In a recent publication Lau et al (2011) demonstrated, how cause-specific and subdistribution hazard rates and consequently hazard ratios can be derived from one mixture model. They proposed to use a flexible parametric mixture model in order to assess time-dependencies of cause-specific and subdistribution hazard ratios. In a data application presented in their article, Lau et al used a three-parameter generalized gamma distribution to model the conditional hazard rates in a mixture model. As the generalized gamma mixture model was found to be numerically unstable in a simulation study (results not shown), an alternative approach for flexible mixture model estimation using penalized B-spline basis function is proposed in this work (Section 5). The approach is based on an estimation procedure for the hazard rate, presented by Rosenberg (1995) for a standard survival model with one possible endpoint, using B-spline functions. The method was adapted for the mixture model approach including a penalty term to obtain smooth and numerical stable estimates for the conditional hazard rates. The newly proposed approach was compared to parametric mixture models assuming the conditional event times to follow Weibull or generalized gamma distributions in a setting with a binary covariate and two possible types of event, regarding their abilities to reflect given cause-specific and subdistribution hazards and their numerical properties. Competing risks data were generated following predefined cause-specific or subdistribution hazards using algorithms provided by Beyersmann et al (2009), which were described and investigated in Section 6 of this work and partly published in Haller and Ulm (2013). The simulation study revealed, that the newly developed spline approach was able to reflect true underlying cause-specific and subdistribution hazards and hazard ratios, even when these were time-dependent, but variability of the obtained results was high, especially for late timepoints with low amount of information left, when the smoothing parameter was chosen to be low in order to allow very flexible estimation of the conditional hazard rates. Using a higher value for the smoothing parameter lead to less variable results, but also to estimates for the conditional hazard rates and consequently for the cause-specific and subdistribution hazard rates, that tended towards a straight line resulting in slightly biased estimates, when shapes of the true hazard rates were time-dependent. Parametric mixture models, that only assessed group effects

on the location parameter, provided biased estimates for cause-specific hazard rates and hazard ratios, when shapes of the underlying hazard rates differed between the groups. A generalized gamma mixture model, allowing location, shape, and scale parameters to vary between the groups, was able to reflect the general shape of the underlying hazard rates, but estimates were biased and very variable for early timepoints. Maximum likelihood estimates could not be determined for all generated datasets, as the numerical maximization procedure did not converge in some cases. In scenarios with a high proportion of censored observations, this was the case for up to almost one third of the generated datasets, when a generalized gamma mixture model, that allowed different shape and scale parameters for both groups, was applied, whereas maximum likelihood estimates could be determined for all generated datasets, when the spline approach with a smoothing parameter of μ=1 or μ=100 was considered. In summary, the new spline-based approach with an adequate choice for the smoothing parameter appears to be superior to the parametric mixture models for deviation of cause-specific and subdistribution hazard rates and hazard ratios, when the true cause-specific and subdistribution hazard rates are time-dependent. A wider discussion on the results of the simulation study can be found in Section 7.5.
For convenience and ease of notation, the new approach was presented for a comparison of two groups in the presence of two possible types of event. The approach can be adapted to more complex settings, but additional assumptions have to be made and model notation has to be adapted. For the simulation study a fixed number of five interior knots for definition of the set of basis functions was used and three different values for the smoothing parameter were investigated. So no recommendations for number and placing of interior knots and the choice of the smoothing parameter can be given yet, as this will be subject to further investigations. One possible strategy for the choice of the smoothing parameter can be the determination of an optimal parameter by the means of a cross validation procedure (see e.g. Hastie et al, 2009).

9.2 Conclusion

This works describes different methods for analysis of competing risks data. The adequate choice of the methods to use is still under discussion, but in recent years most authors argued for modelling the whole competing risks process, which is naturally defined by the cause-specific hazard rates (Beyersmann et al, 2012; Andersen and Keiding, 2012; Koller et al, 2012). Nevertheless, other approaches might provide additional information in certain applications or might be the better fit to answer specific research questions. The subdistribution hazards regression allows a direct translation of the covariate effects on the hazard rate to an effect on the event probability, which appears to be much more intuitive for applicants and readers not familiar with the concept of hazard rates. Mixture models, although conditioning on future events and therefore not allowing adequate prediction of an individual's event time, provide marginal event type probabilities. Application of the vertical modelling approach gives estimates of the marginal event time distribution and of event type probabilities for given event times. Cause-specific and subdistribution hazard rates can be derived from a mixture model, allowing to assess estimates for both quantities from one model as described by Lau et al (2011). A newly proposed spline approach

performed better than parametric mixture models in a simulation study.
One major task remains the transfer of available methods for the analysis of competing risks data to the medical community, in order to avoid misinterpretation of study data, possibly leading to erroneous therapy decisions or risk stratifications, due to inadequate application of statistical methods in the presence of competing risks.

Appendix A

Simulation with predefined subdistribution hazards

The R-code for simulation of competing risks data with predefined subdistribution hazards for the multiple regression model shown in Section 6.3.4 is presented.

```
# Simulation of a multiple regression model - Example 5

# Required libraries
library(numDeriv) # Compute derivatives
library(cmprsk)   # Estimate subdistribution hazard ratios
                  # assuming proportionality

# Defining baseline hazards for the event of interest
gamma1_0 <- function(t) 0.001 * exp(-0.001*t/log(1.3))
lambda1_0 <- function(t) 0.001
# Calculating the cause-specific baseline hazard for
# the competing event
lambda2_0 <- function(t)
gamma1_0(t) - lambda1_0(t) - grad(gamma1_0,t)/gamma1_0(t) +
                       grad(lambda1_0,t)/lambda1_0(t)

# Defining the regression coefficients (=log-hazard ratios)
xi <- c(log(1.15),log(0.9),log(2))

# Hazards as functions of time, regression coefficients and
# covariate values
gamma1_x <- function(t,xi,X) gamma1_0(t)*exp(sum(xi*X))
lambda1_x <- function(t,X)
    lambda1_0(t)*exp(xi[1]*X[1]*exp(-0.0005*t) +
       xi[2]*X[2]*exp(-0.0005*t) + xi[3]*X[3]*exp(-0.0005*t))

# Calculating the cause-specific hazard ratio for the event of
# interest dependening on the covariates
lambda2_x <- function(t,xi,X)
   gamma1_x(t,xi,X) - lambda1_x(t,X) -
```

```
      grad(gamma1_x,t,xi=xi,X=X)/gamma1_x(t,xi,X) +
         grad(lambda1_x,t,X=X)/lambda1_x(t,X)

# Alternatively - to save computation time -
# the cause-specific hazard # for the
# competing event can be determined analytically
# lambda2_x <- function(t,xi,X)
#   0.001*exp(-0.001*t/log(1.3)+sum(xi*X)) -
#      0.001*exp(sum(xi*X*exp(-0.0005*t))) + 0.001/log(1.3) -
#         sum(xi*X*0.0005*exp(-0.0005*t))


# Number of simulation runs
Runs <- 4000
# Number of subjects in each run
n <- 1000
# Matrix to save results
CRR <- matrix(nrow=Runs,ncol=3)

# Loop for repeated runs
for(RR in 1:Runs)
{
# Random sampling of covariates from the defined distributions
# X2 is restricted to values between 0 and 10
  X1 <- runif(n,0,3)
  X2 <- pmax(pmin(rnorm(n,5,1),10),0)
  X3 <- sample(0:1,n,repl=T)
  XX <- cbind(X1,X2,X3)

# Vectors to save results within a loop
  Evstat <- rep(0,n)
  Evtime <- c()

  # determine event time and type for each individual
  for(j in 1:n)
  {
    Ti <- 0     # time variable

    while(Evstat[j]==0) # repeat until an event was observed
    {
      Ti <- Ti+1
      # Probability of an event for individual j at time Ti
      Prob <- lambda1_x(t=Ti,X=XX[j,]) +
                  lambda2_x(t=Ti,xi=xi,X=XX[j,])
      # Determine, if any event happened at Ti
      Event <- sample(0:1,1,prob=c(1-Prob,Prob))
  # if an event happened, determine which type
      if(Event==1)
        Evstat[j] <- sample(1:2,1,prob=c(
```

```
          lambda1_x(t=Ti,X=XX[j,]),
          lambda2_x(t=Ti,xi=xi,X=XX[j,])))
    }
    Evtime[j] <- Ti
  }

  # Transform generated data to a data frame
  dat <- data.frame(Time=Evtime,Stat=Evstat,X1=X1,X2=X2,X3=X3)

  # Perform Fine and Gray regression and save
  # the regression coefficients
  CRR[RR,] <- crr(dat$Time,dat$Stat,
            cov1=cbind(dat$X1,dat$X2,dat$X3))$coef
  # Calculate means of current
  # subdistribution log-hazard ratios
  MEAN <- apply(CRR,2,mean,na.rm=T)
  # Print current proceeding and
  # mean of subdistribution hazard
  # ratios over all loops performed
  print(paste("Run: ",RR,"  -  Means: ",round(MEAN[1],3)," / ",
                round(MEAN[2],3)," / ",round(MEAN[3],3)))
}
```

Appendix B

Appendix to the simulation study

B.1 Further results

B.1.1 Constant cause-specific hazard ratio

Further results for the simulation scenario with a time-dependent monotonous cause-specific hazard ratio, as described in Section 7.3.1 and Section 7.4.1, are shown here. The most important results, namely estimates for the average cause-specific (log-)hazard ratio and illustrations of the estimated cause-specific hazard ratios, are presented in Section 7.4.1. Summaries of estimates of the cause-specific hazards for the event of interest are shown here for both groups as well as summaries of estimates for the subdistribution hazards and hazard ratios.
Expected subdistribution hazard rates were derived from the prespecified cause-specific hazard rates following Equation 3.14 and are illustrated in the according figures by solid grey lines. The expected subdistribution hazard ratio was derived as quotient of the expected subdistribution hazard rates.
Results using the different censoring distributions presented in Section 7.1 are displayed on the following pages:

- Low amount of censored observations - pages 126 to 128
- Moderate amount of censored observations - pages 129 to 131
- High amount of censored observations - pages 132 to 134

Scenario I - Low Censoring

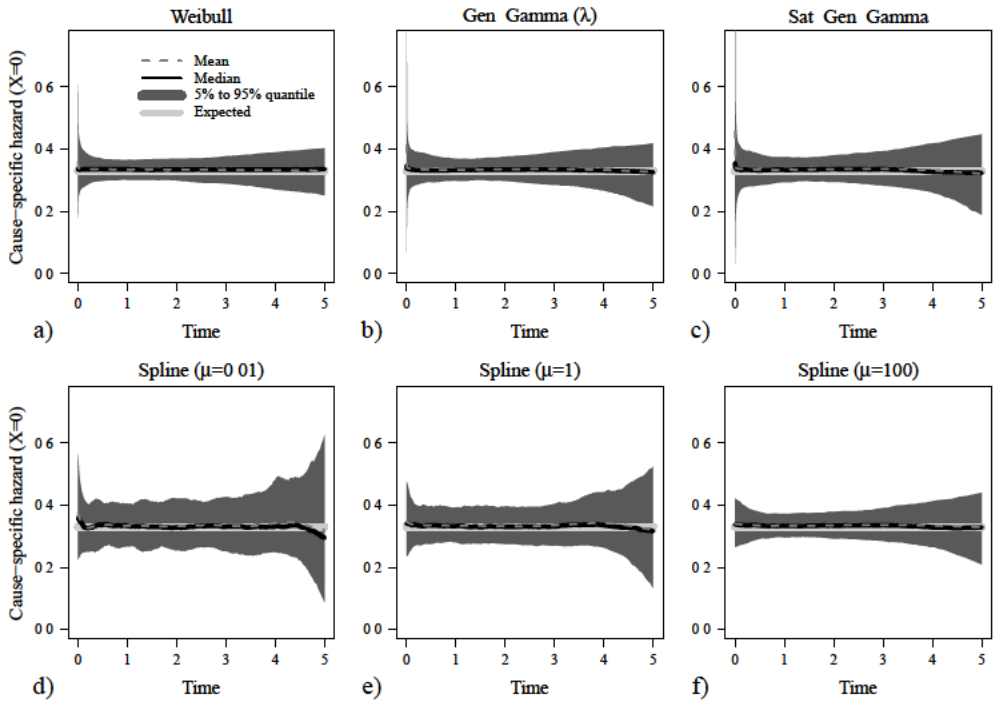

Figure B.1: Scenario I - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

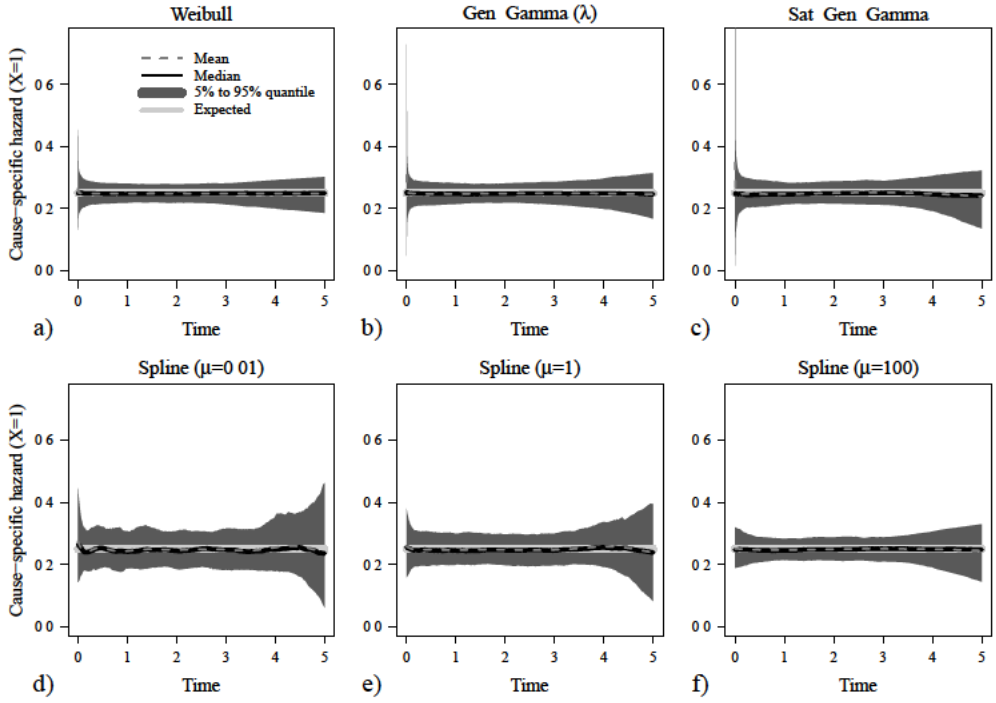

Figure B.2: Scenario I - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

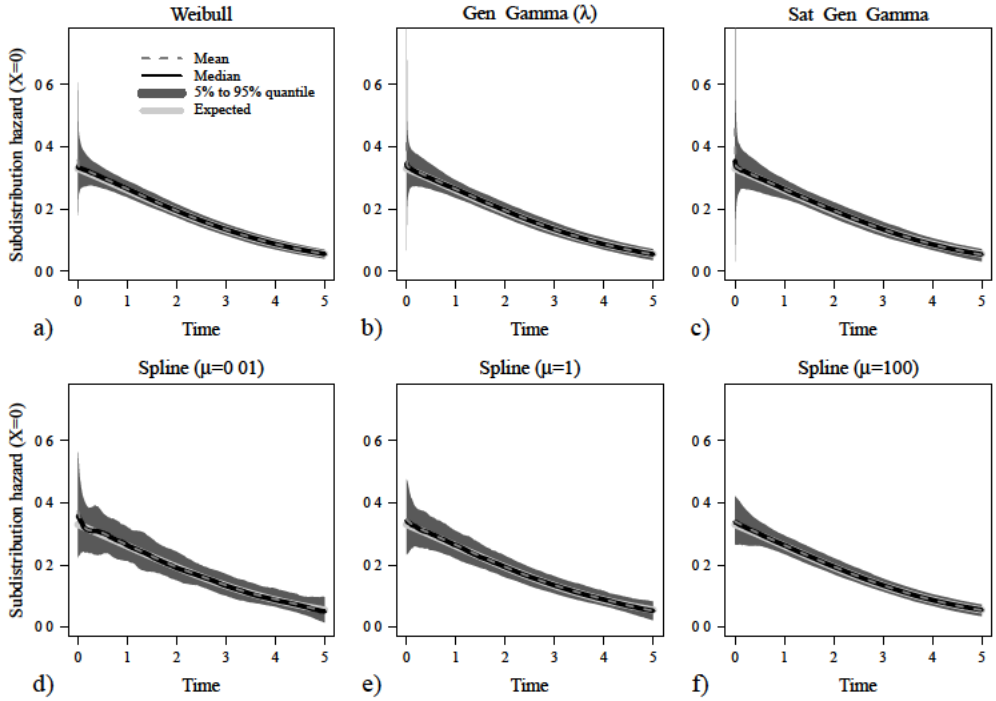

Figure B.3: Scenario I - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0).

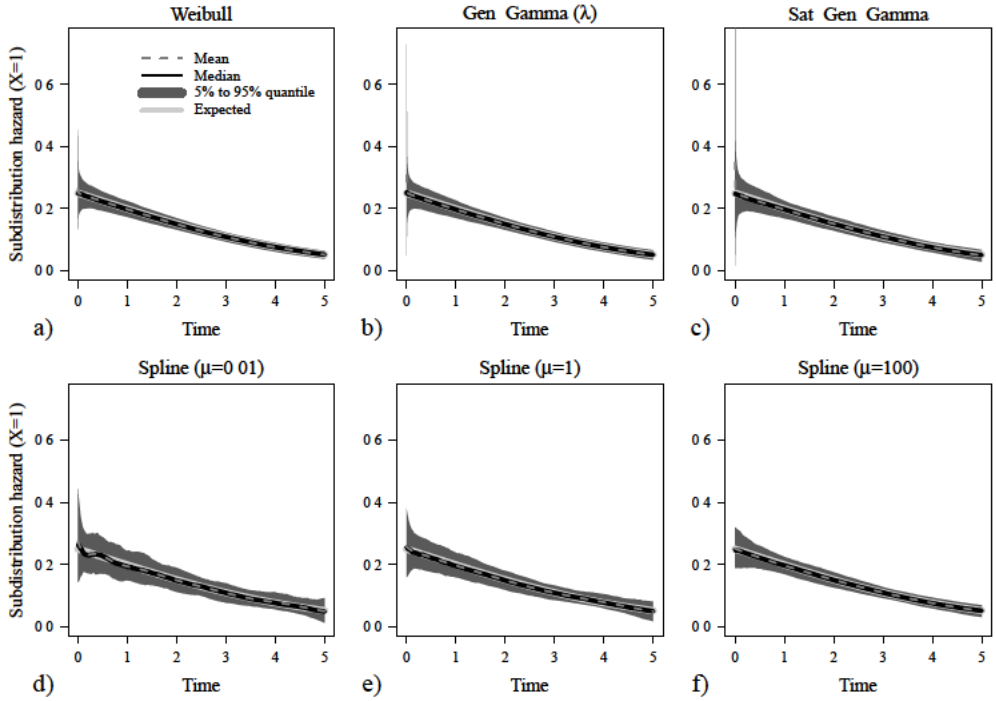

Figure B.4: Scenario I - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

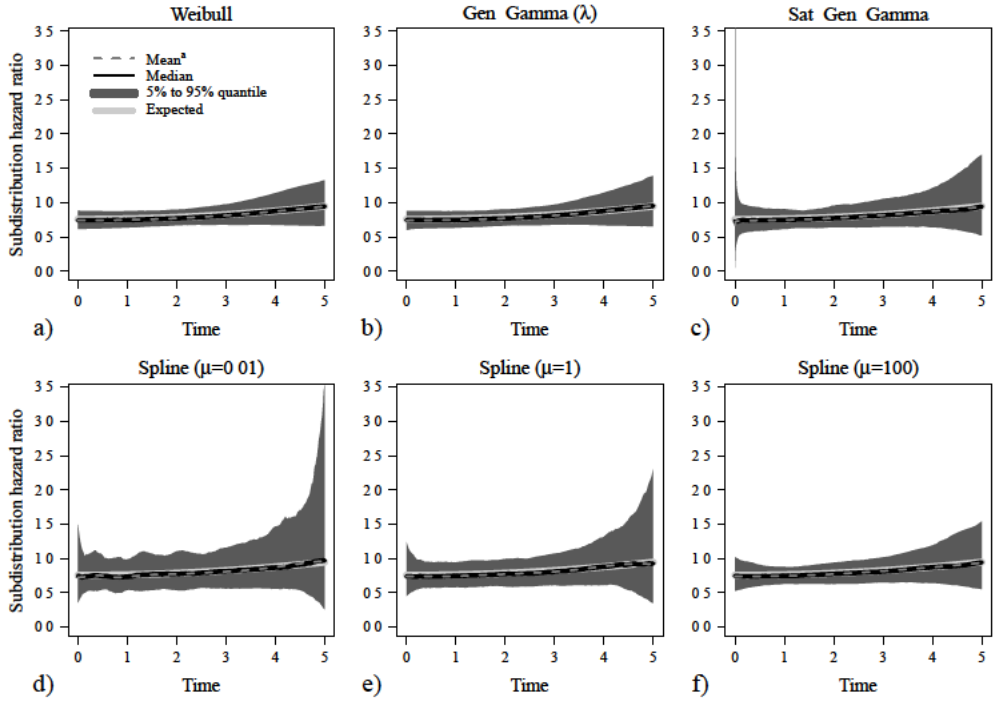

Figure B.5: Scenario I - low censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario I - Moderate Censoring

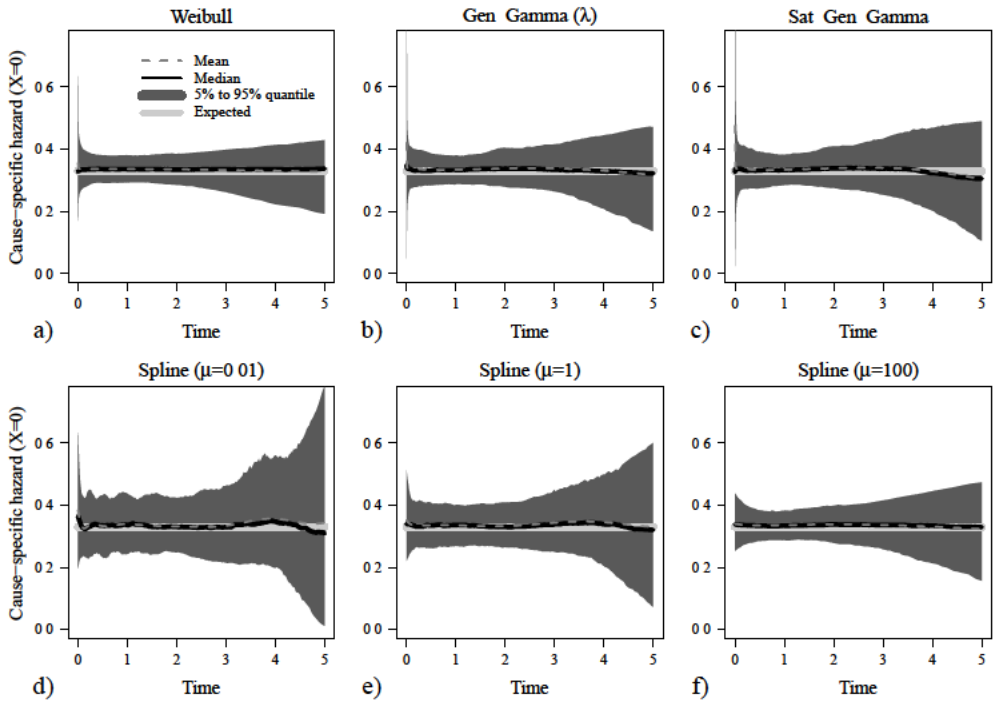

Figure B.6: Scenario I - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

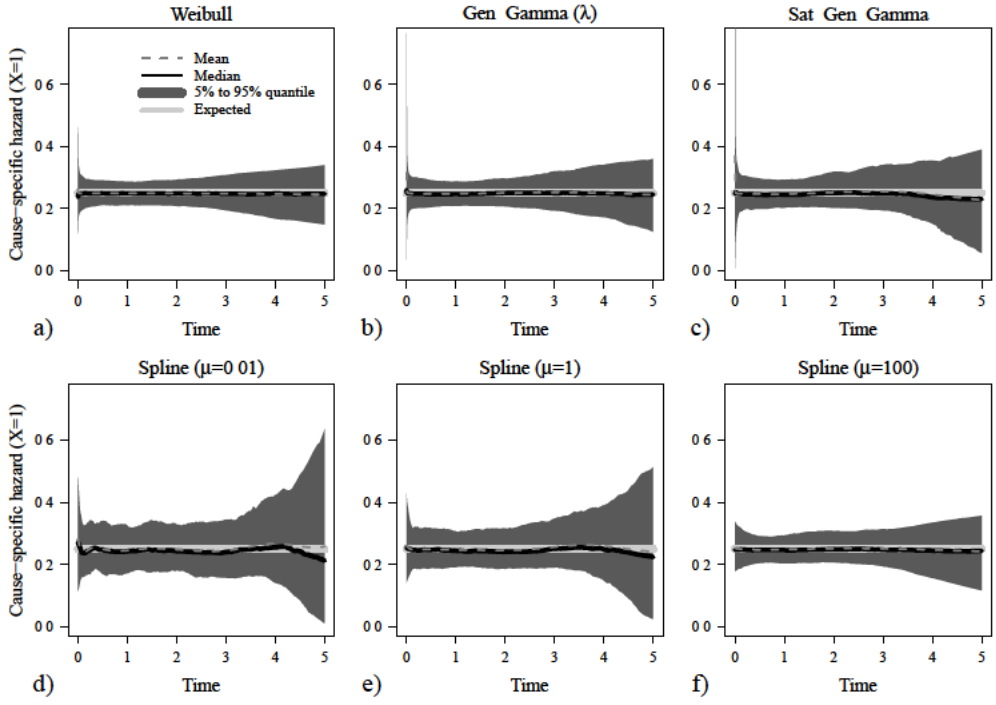

Figure B.7: Scenario I - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

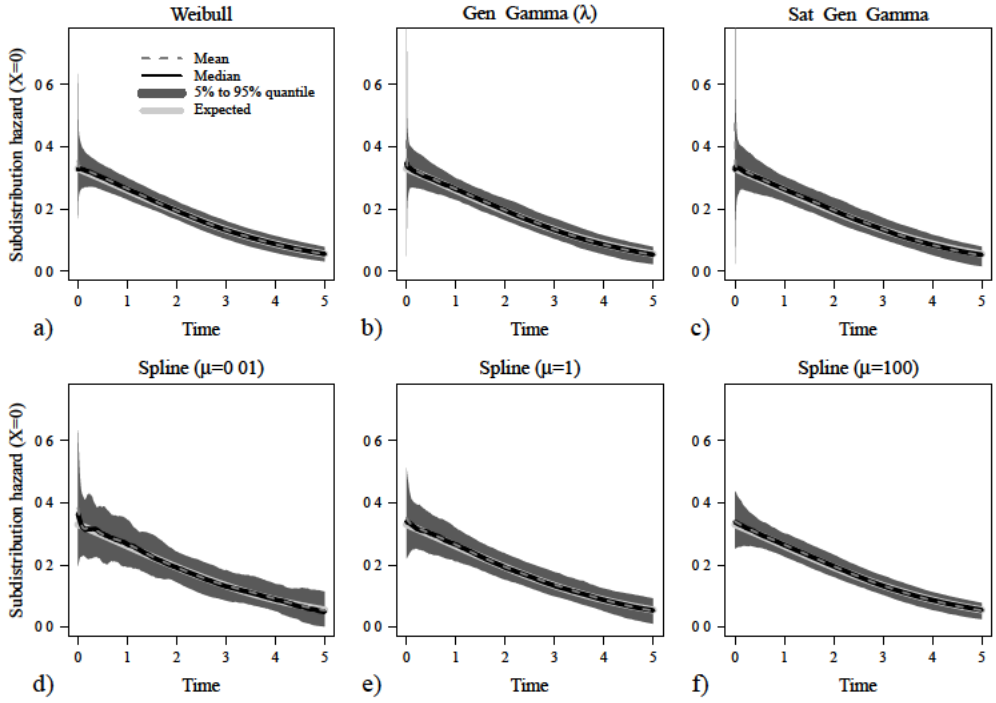

Figure B.8: Scenario I - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

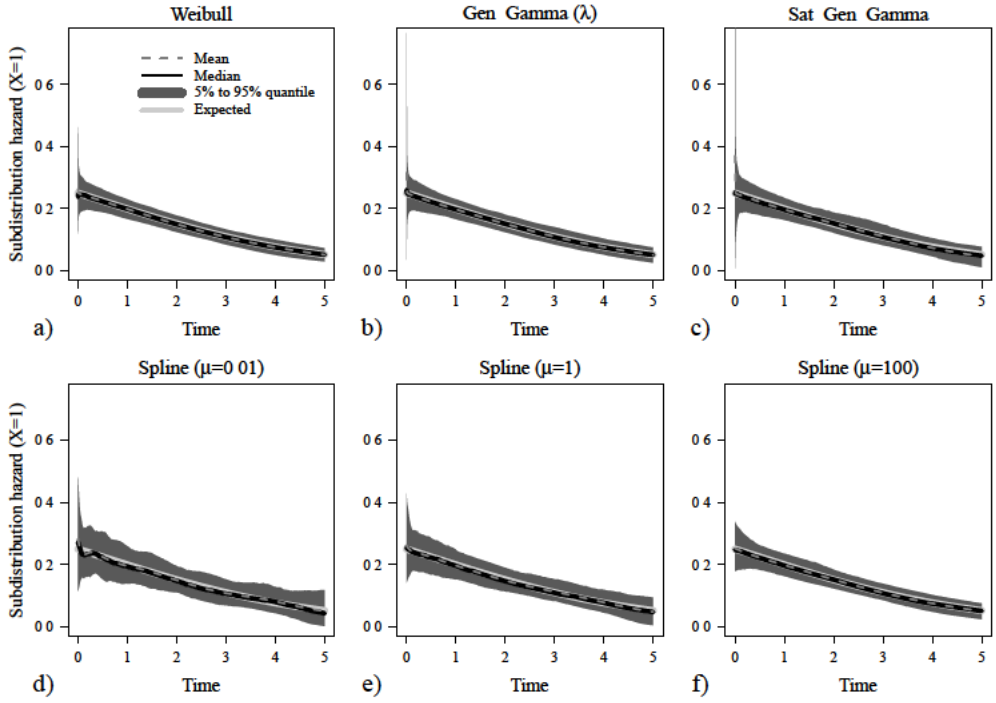

Figure B.9: Scenario I - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

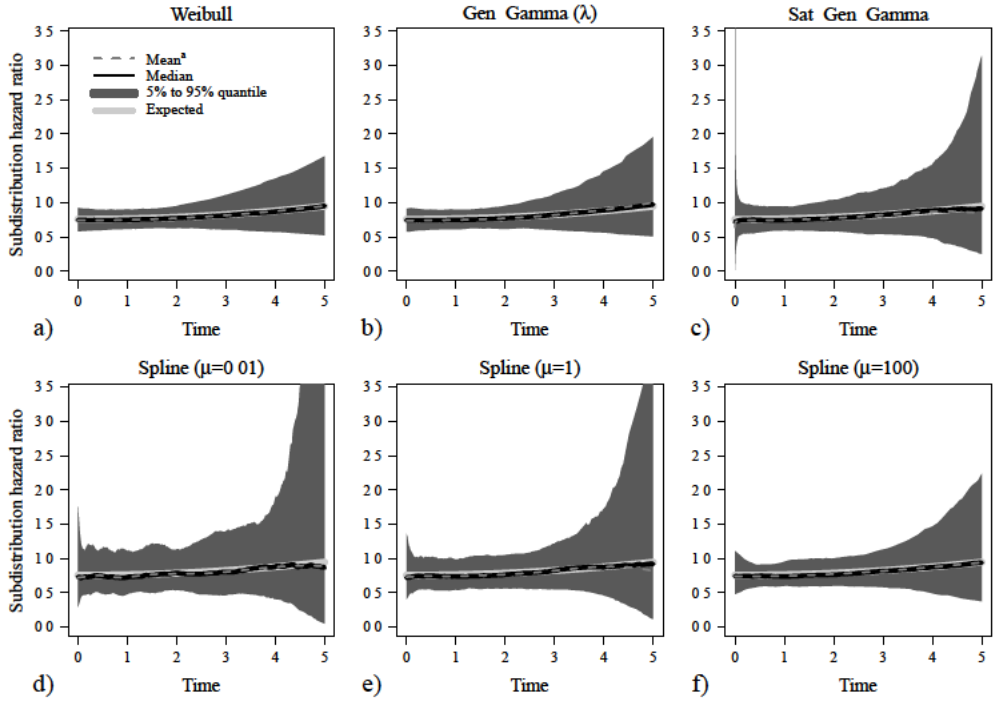

Figure B.10: Scenario I - moderate censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario I - High Censoring

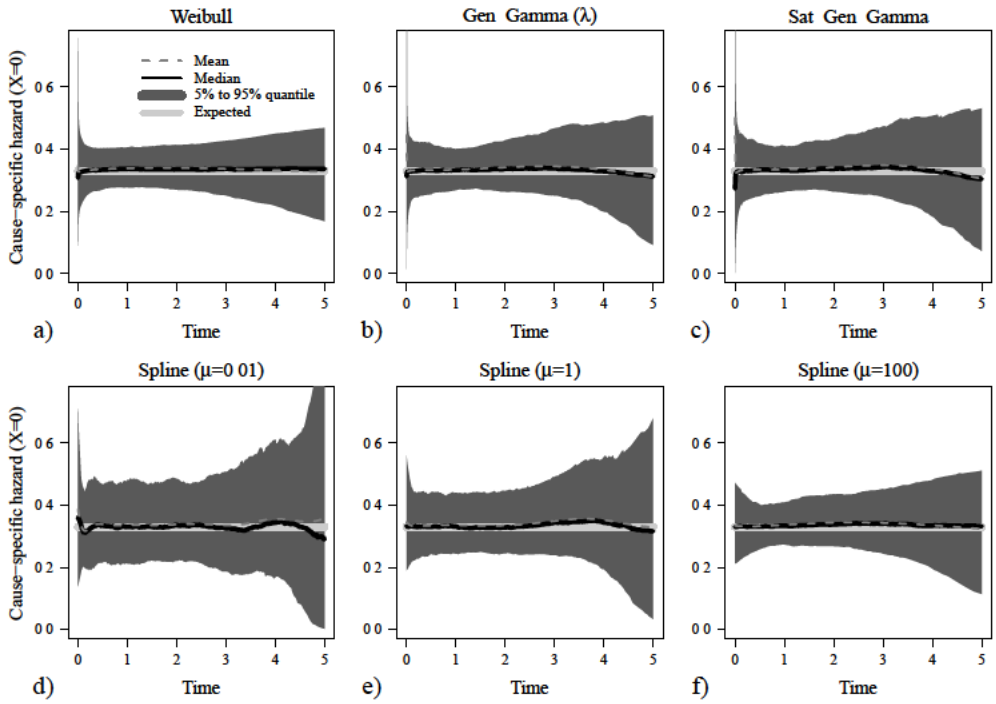

Figure B.11: Scenario I - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

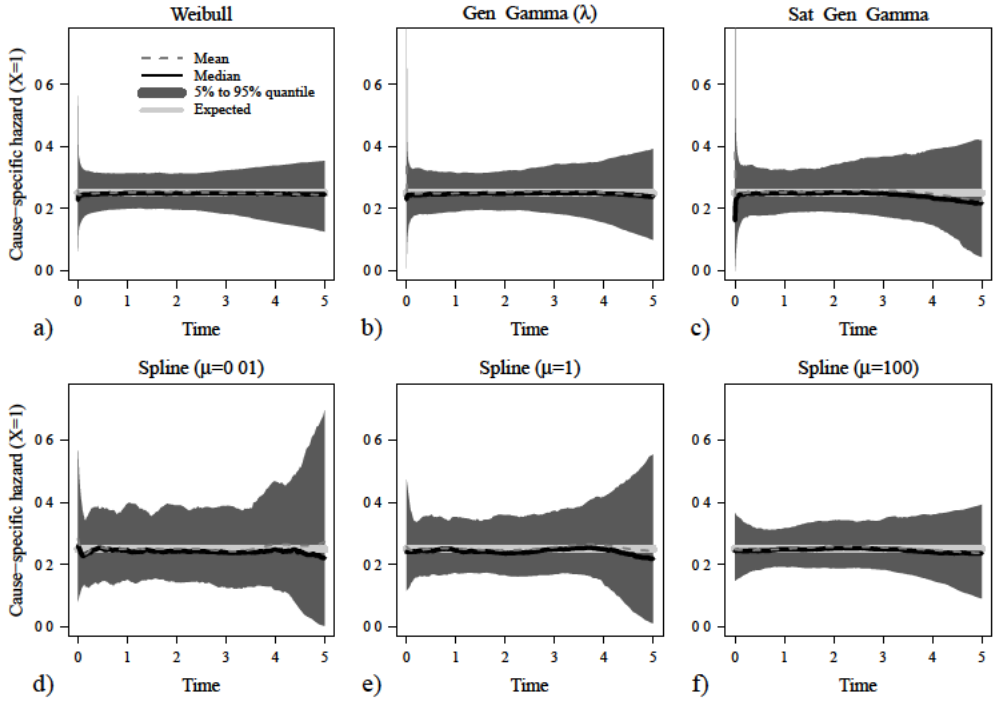

Figure B.12: Scenario I - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

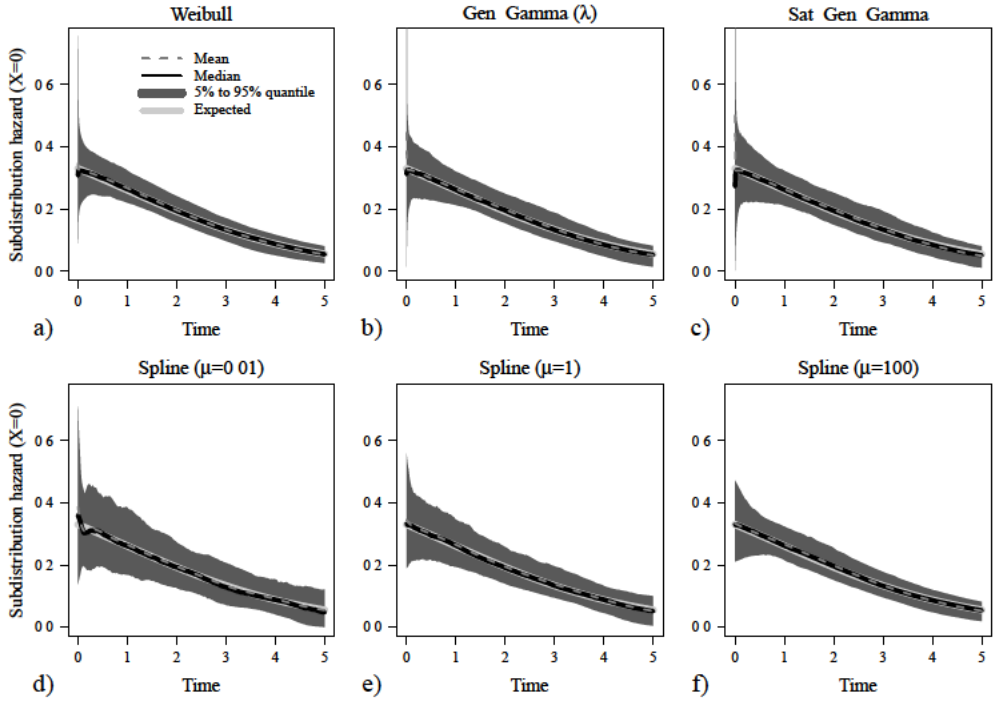

Figure B.13: Scenario I - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

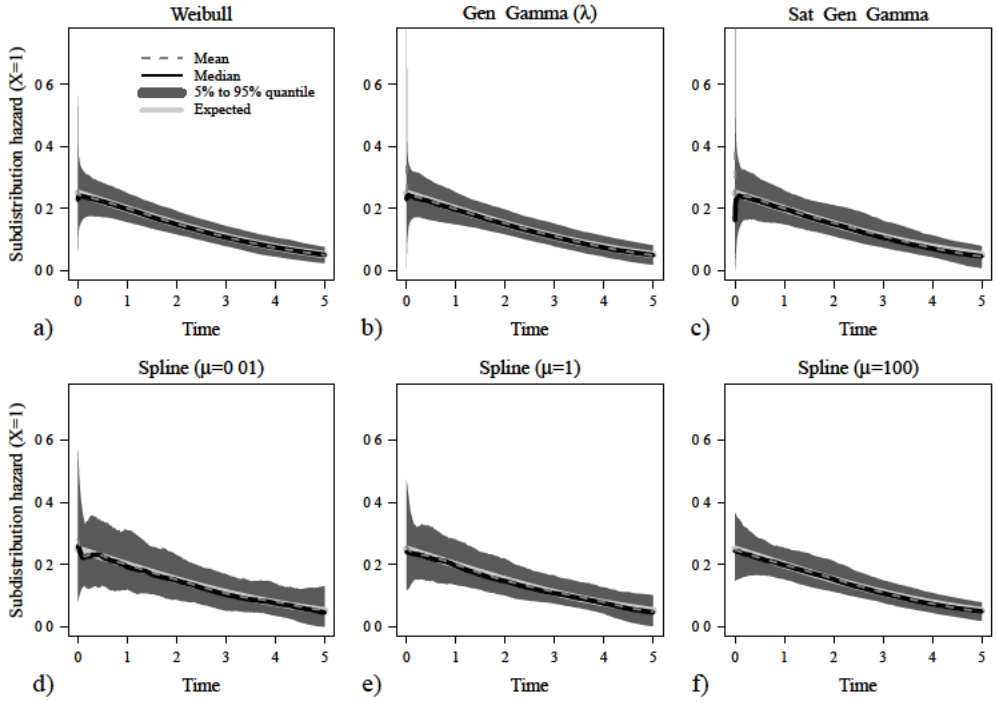

Figure B.14: Scenario I - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

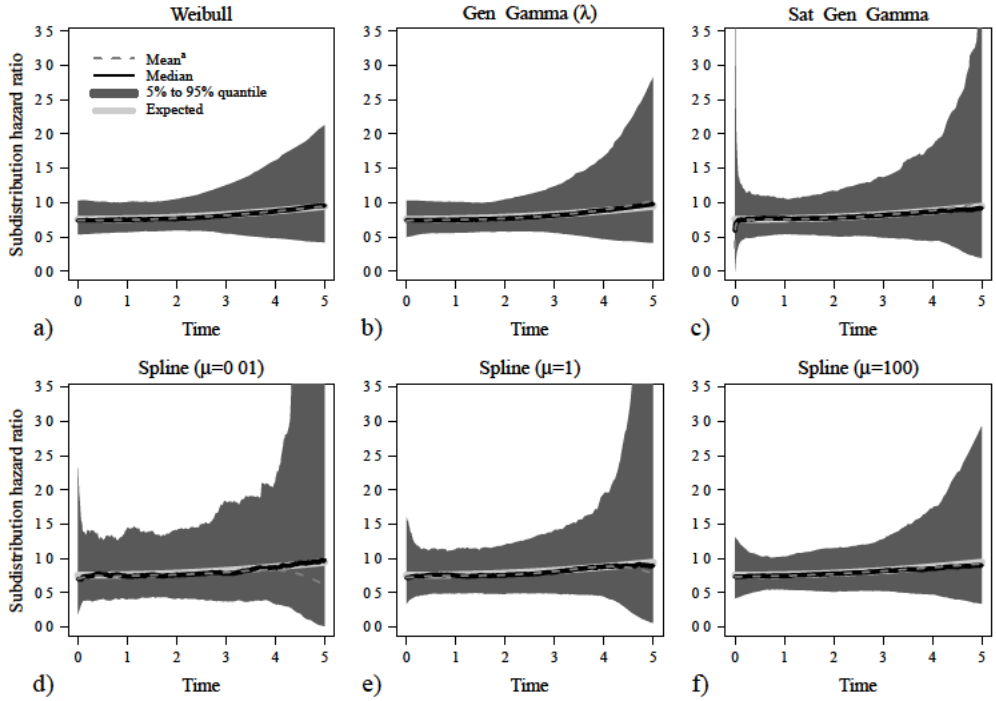

Figure B.15: Scenario I - high censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

B.1.2 Time-dependent monotonous cause-specific hazard ratio

Further results for the simulation scenario with a time-dependent monotonous cause-specific hazard ratio, as described in Section 7.3.2 and Section 7.4.2, are shown here. The most important results, namely illustrations of the estimated cause-specific hazard ratios, are presented in Section 7.4.2. Summaries of estimates for the cause-specific hazards for the event of interest are shown here for both groups as well as summaries of estimates for the subdistribution hazards and hazard ratios.
Expected subdistribution hazard rates were derived from the prespecified cause-specific hazard rates following Equation 3.14 and are illustrated in the according figures by solid grey lines. The expected subdistribution hazard ratio was derived as quotient of the expected subdistribution hazard rates.
Results using the different censoring distributions presented in Section 7.1 are displayed on the following pages:

- Low amount of censored observations - pages 136 to 138
- Moderate amount of censored observations - pages 139 to 141
- High amount of censored observations - pages 142 to 144

Scenario II - Low Censoring

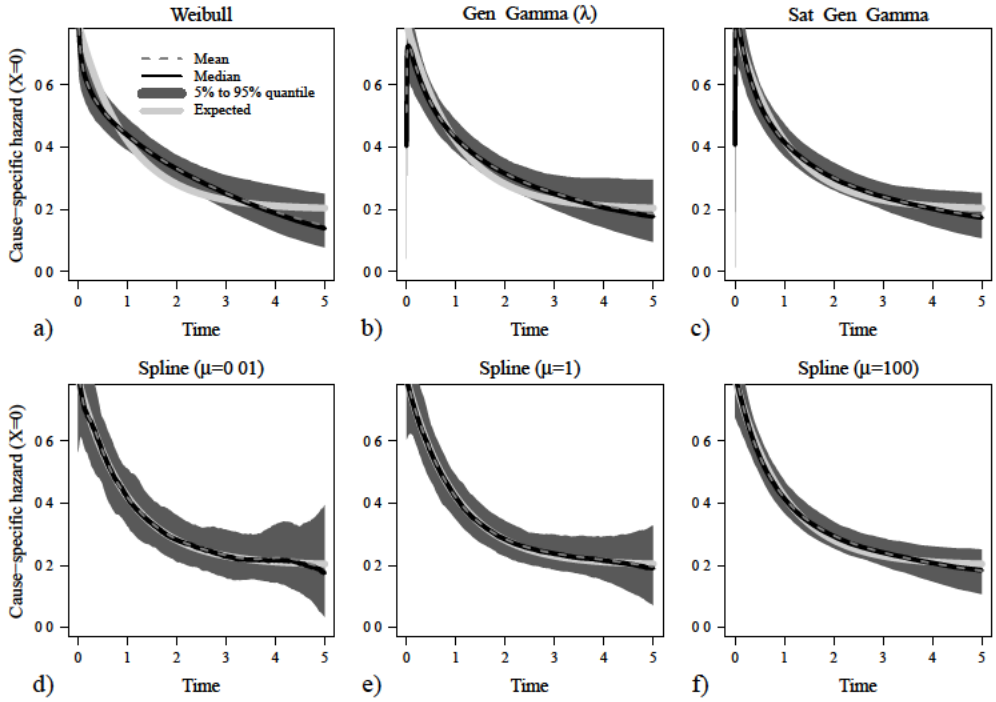

Figure B.16: Scenario II - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

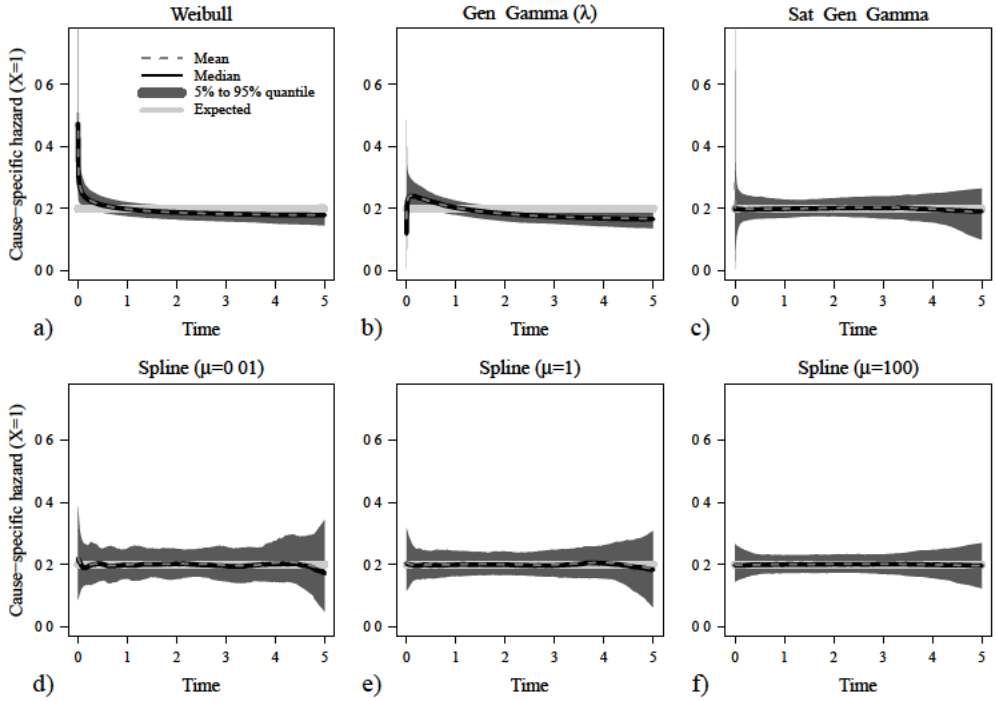

Figure B.17: Scenario II - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

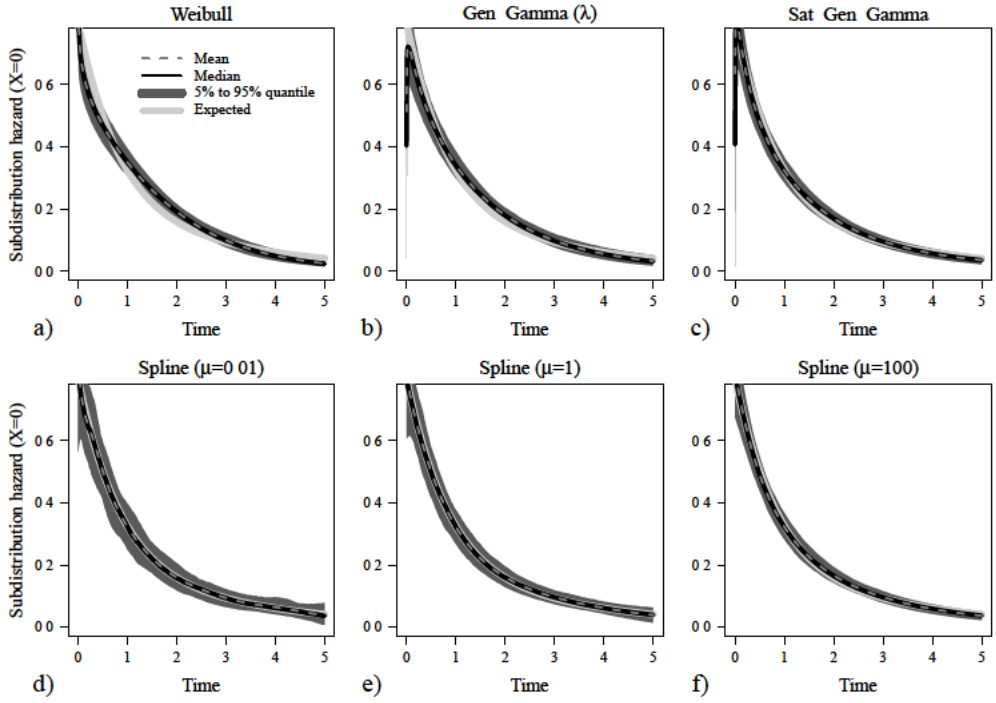

Figure B.18: Scenario II - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

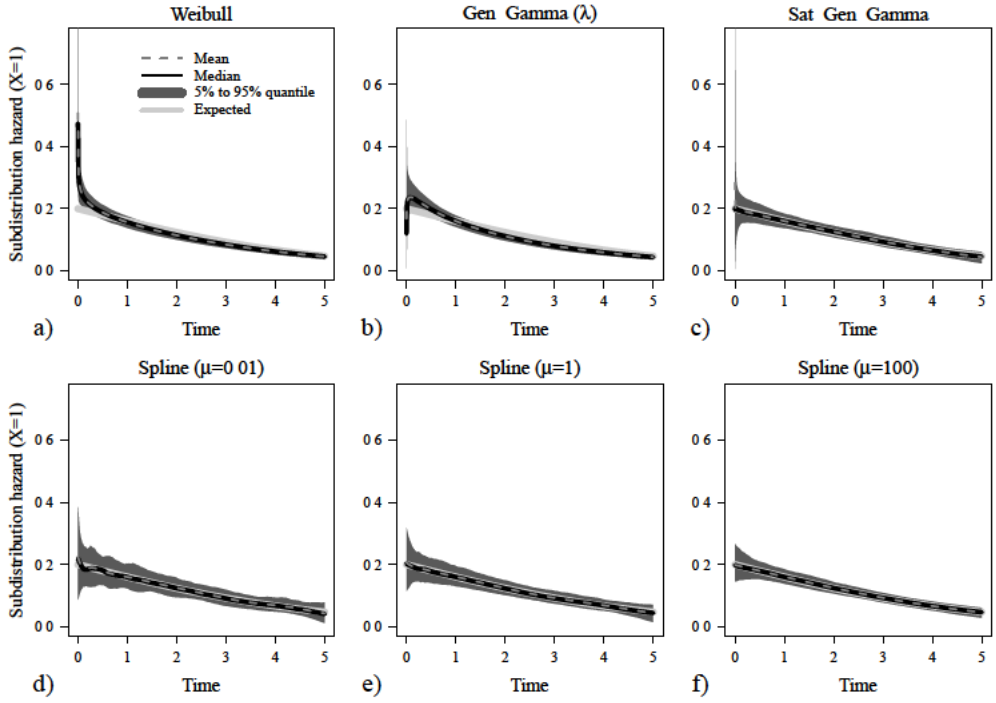

Figure B.19: Scenario II - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

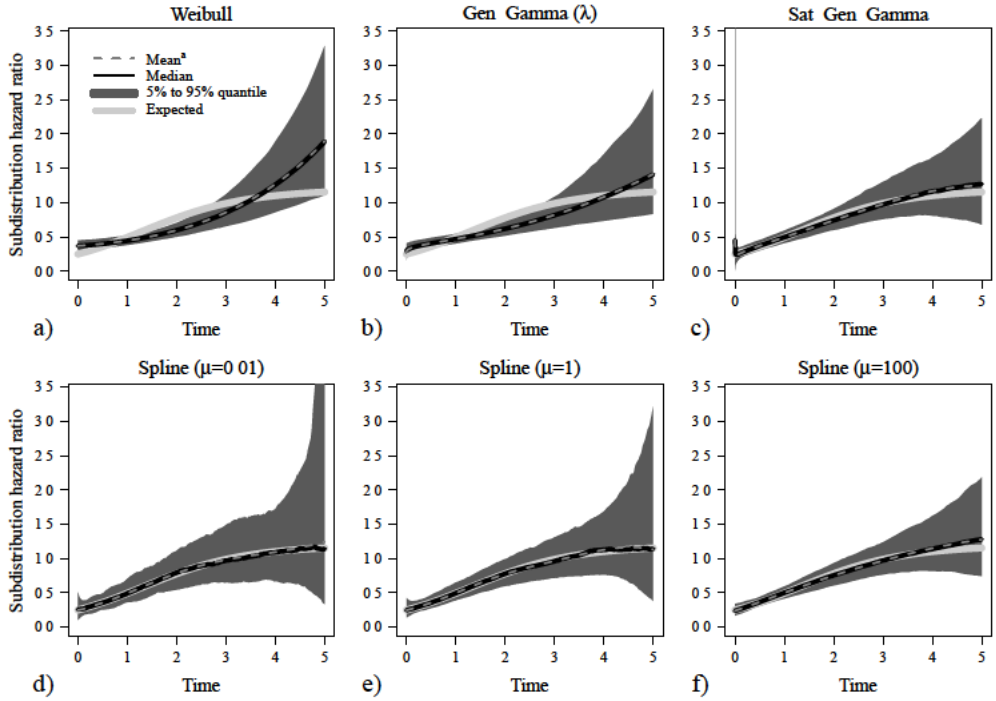

Figure B.20: Scenario II - low censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario II - Moderate Censoring

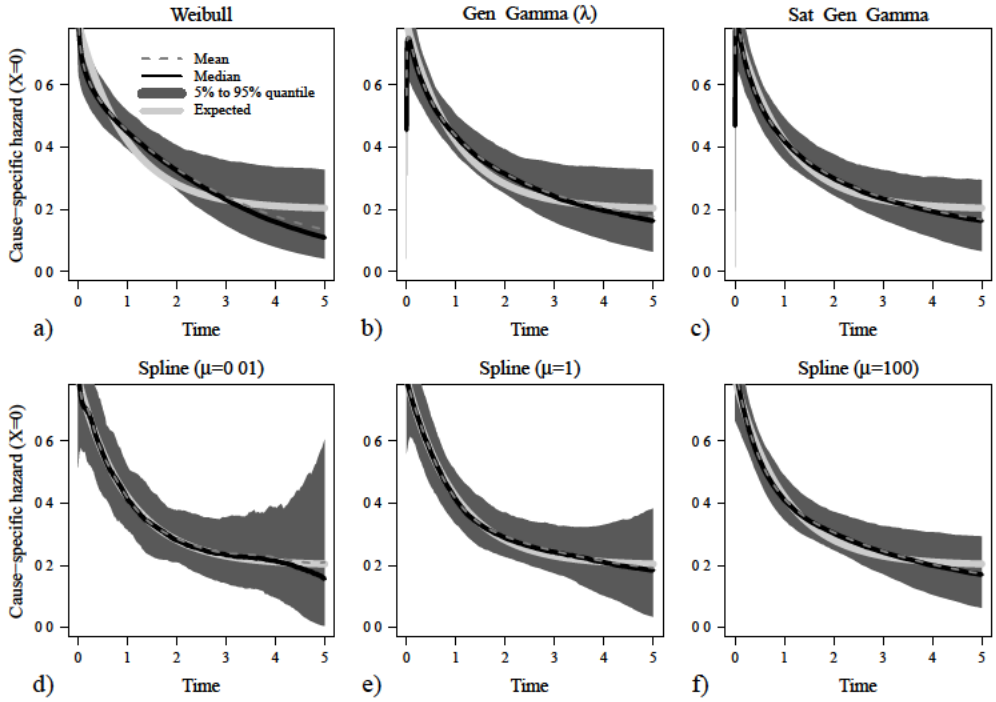

Figure B.21: Scenario II - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

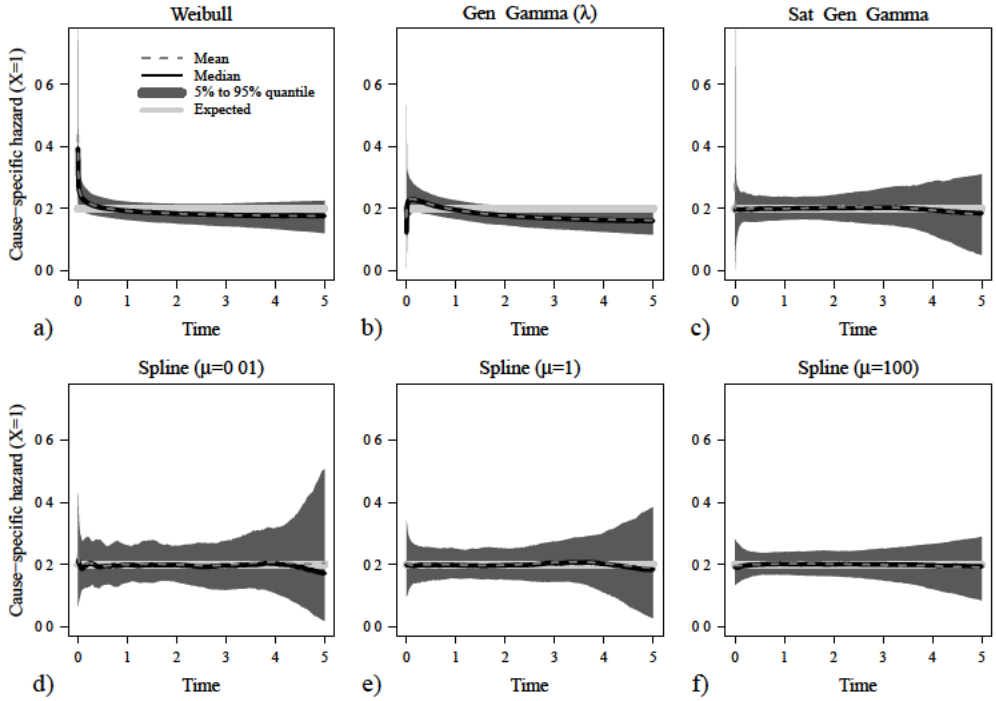

Figure B.22: Scenario II - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

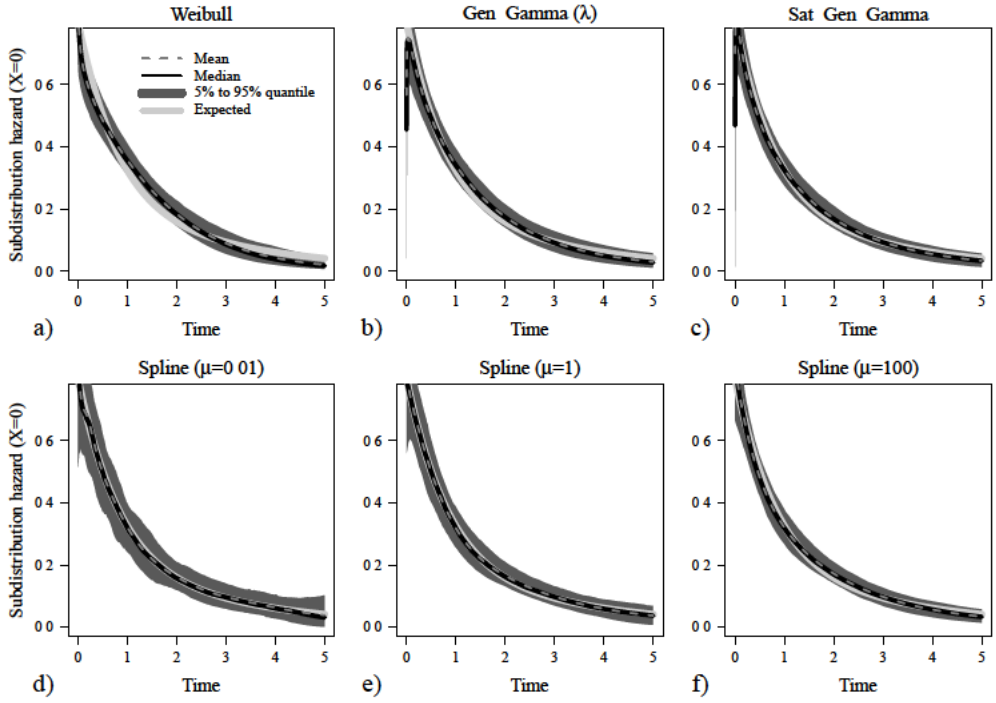

Figure B.23: Scenario II - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

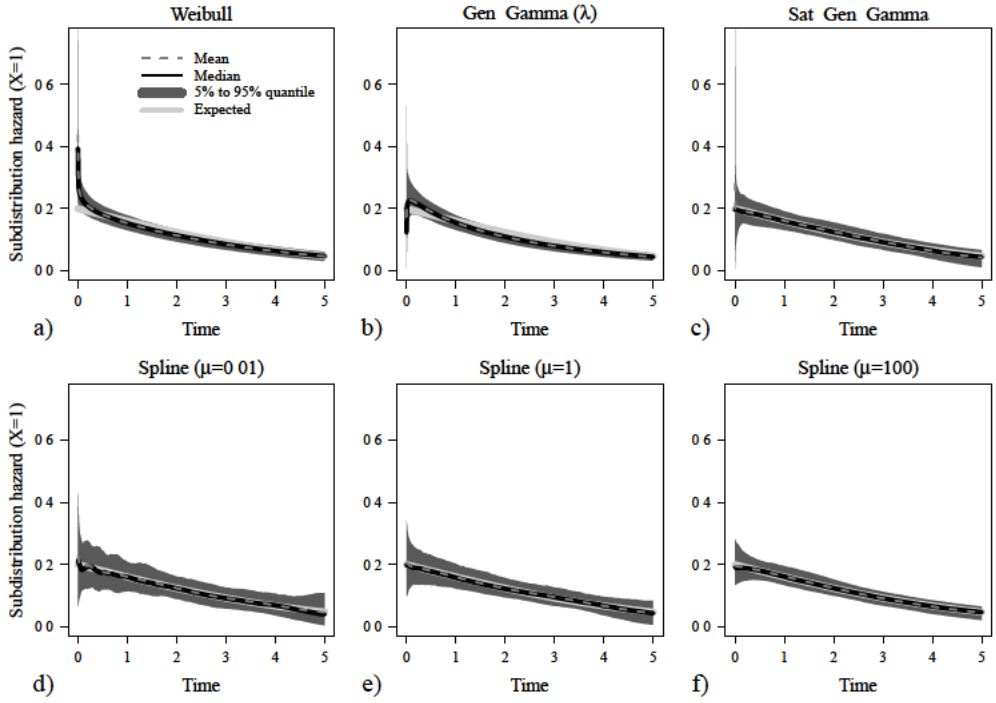

Figure B.24: Scenario II - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

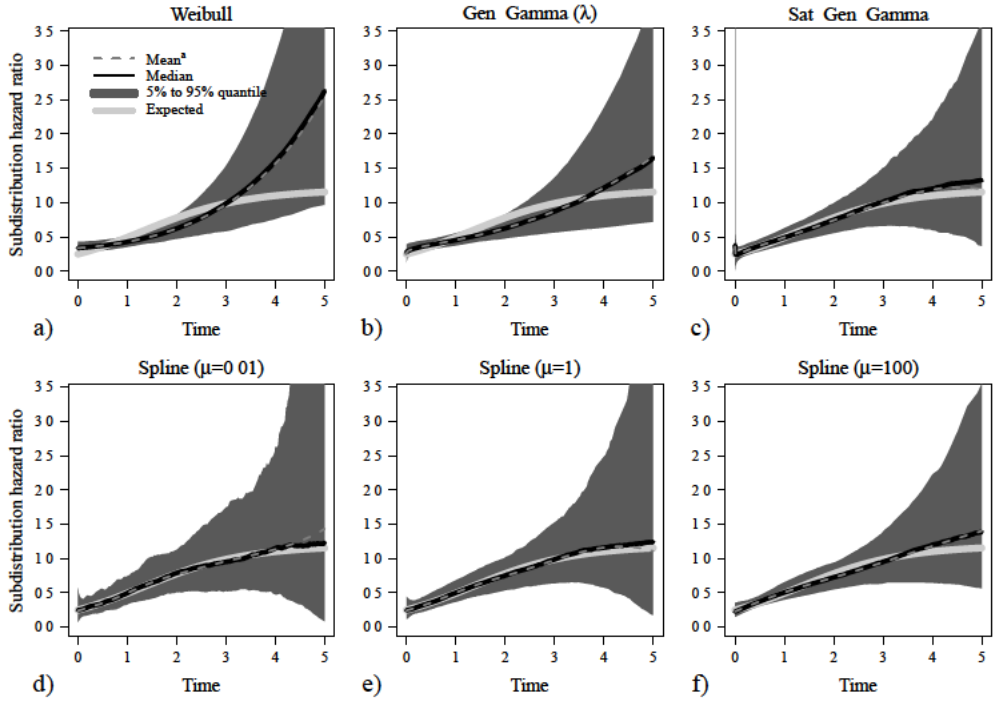

Figure B.25: Scenario II - moderate censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario II - High Censoring

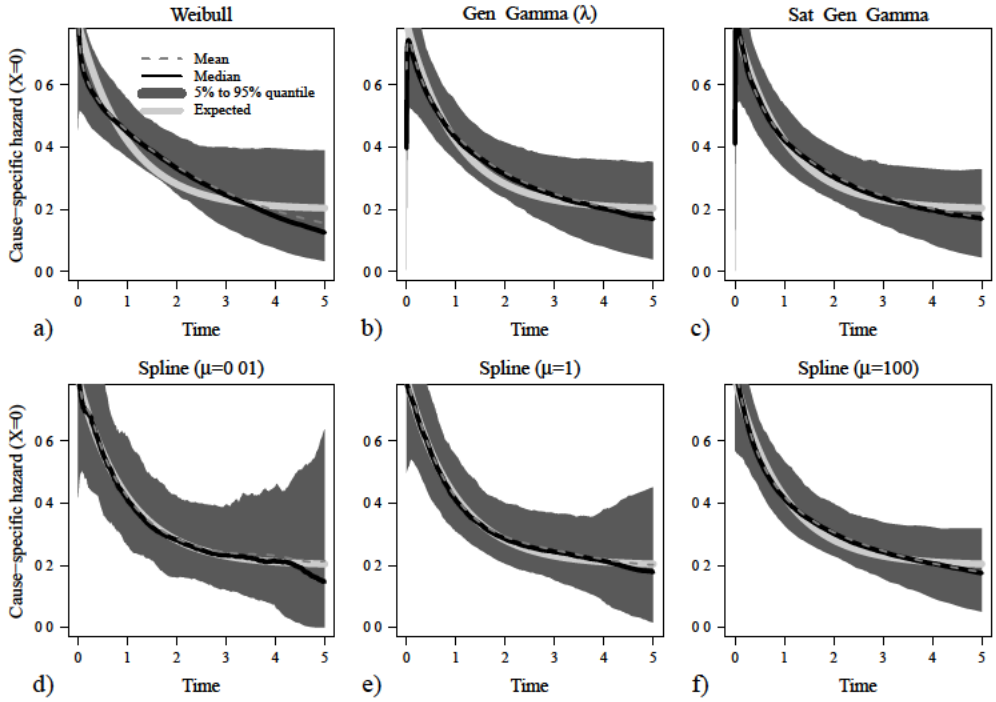

Figure B.26: Scenario II - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

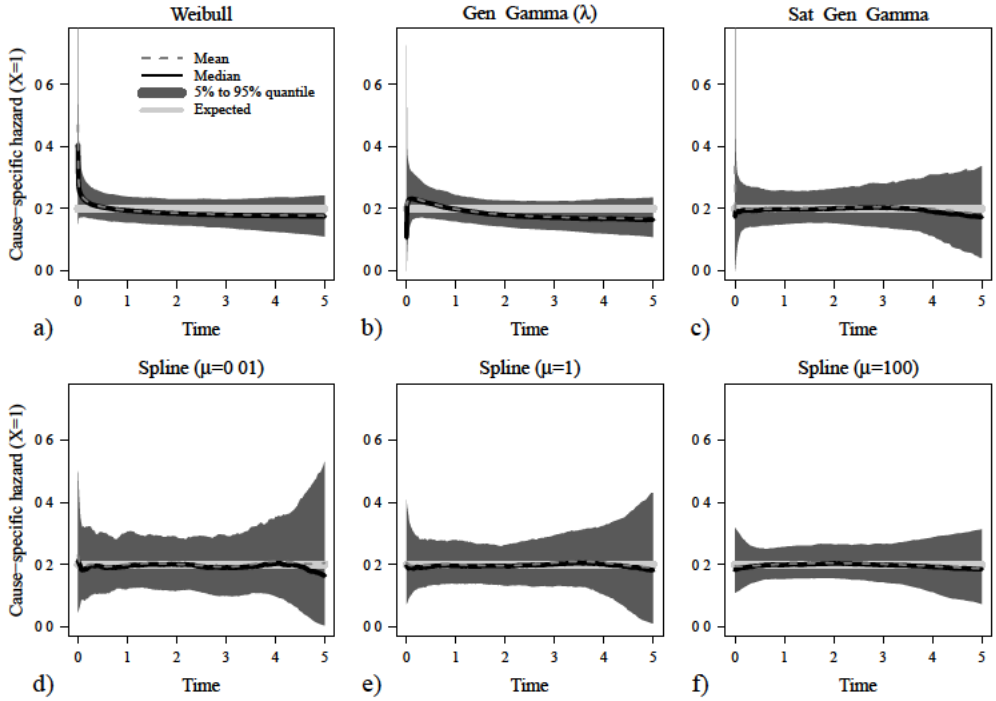

Figure B.27: Scenario II - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

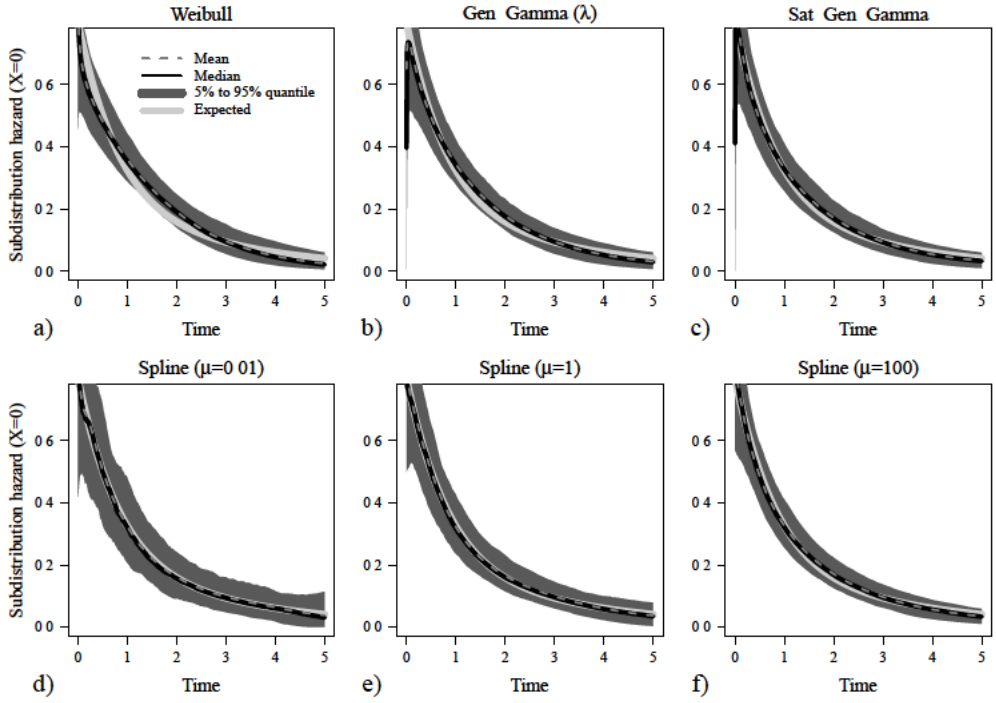

Figure B.28: Scenario II - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

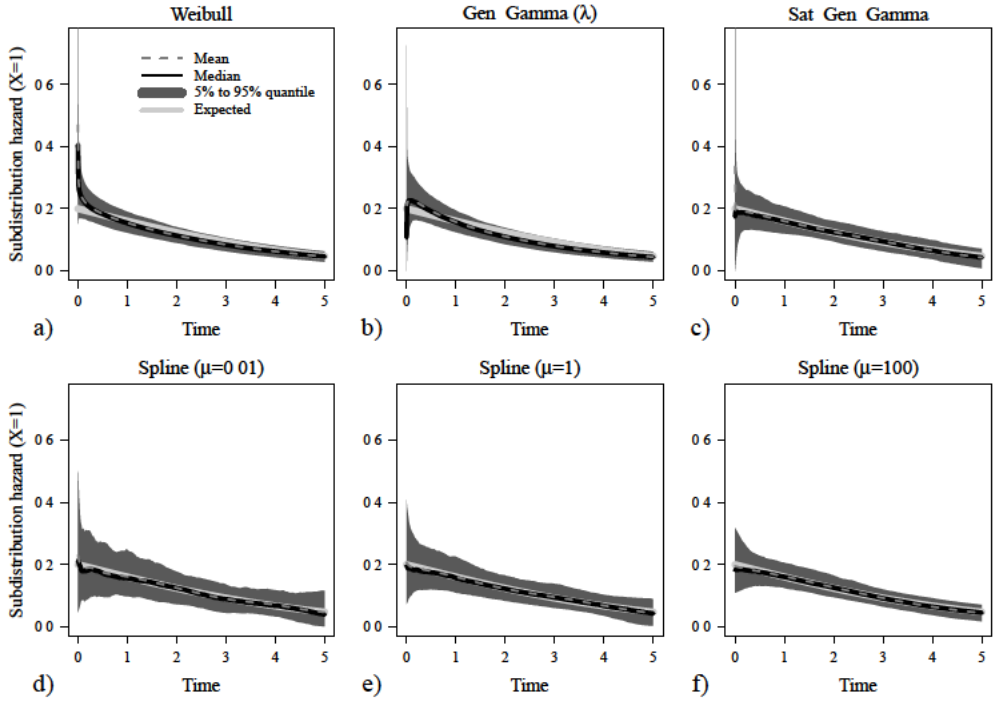

Figure B.29: Scenario II - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

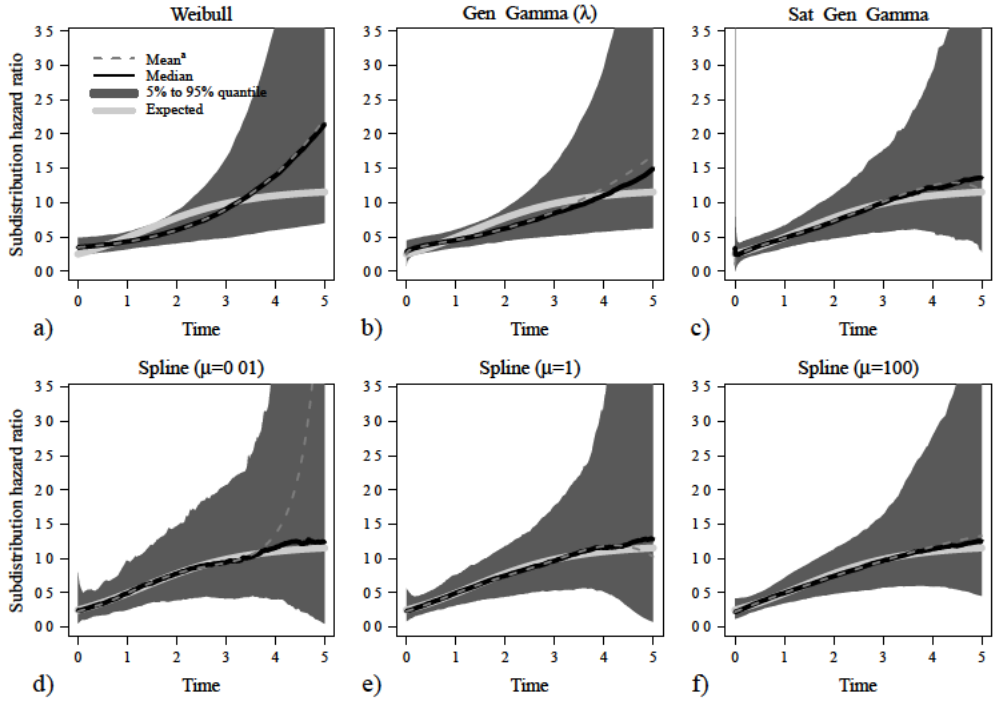

Figure B.30: Scenario II - high censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

B.1.3 Time-dependent non-monotonous cause-specific hazard ratio

Further results for the simulation scenario with a time-dependent non-monotonous cause-specific hazard ratio, as described in Section 7.3.3 and Section 7.4.3, are shown here. The most important results, namely illustrations of the estimated cause-specific hazard ratios, are presented in Section 7.4.3. Summaries of estimates for the cause-specific hazards for the event of interest are shown here for both groups as well as summaries of estimates for the subdistribution hazards and hazard ratios.
Expected subdistribution hazard rates were derived from the prespecified cause-specific hazard rates following Equation 3.14 and are illustrated in the according figures by solid grey lines. The expected subdistribution hazard ratio was derived as quotient of the expected subdistribution hazard rates.
Results using the different censoring distributions presented in Section 7.1 are displayed on the following pages:

- Low amount of censored observations - pages 146 to 148
- Moderate amount of censored observations - pages 149 to 151
- High amount of censored observations - pages 152 to 154

Scenario III - Low Censoring

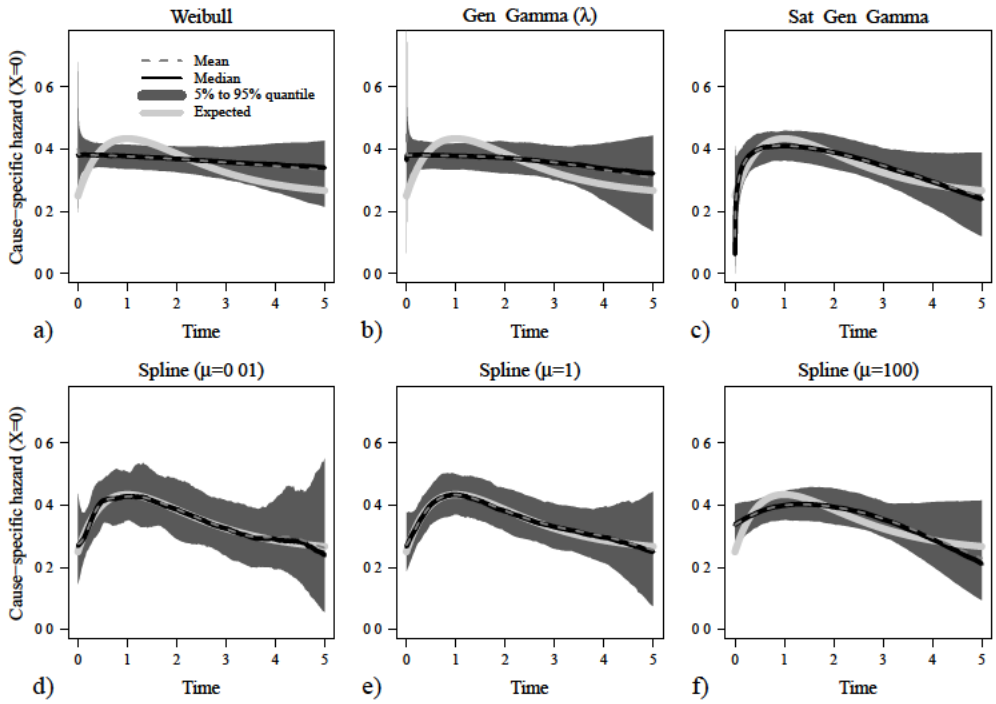

Figure B.31: Scenario III - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

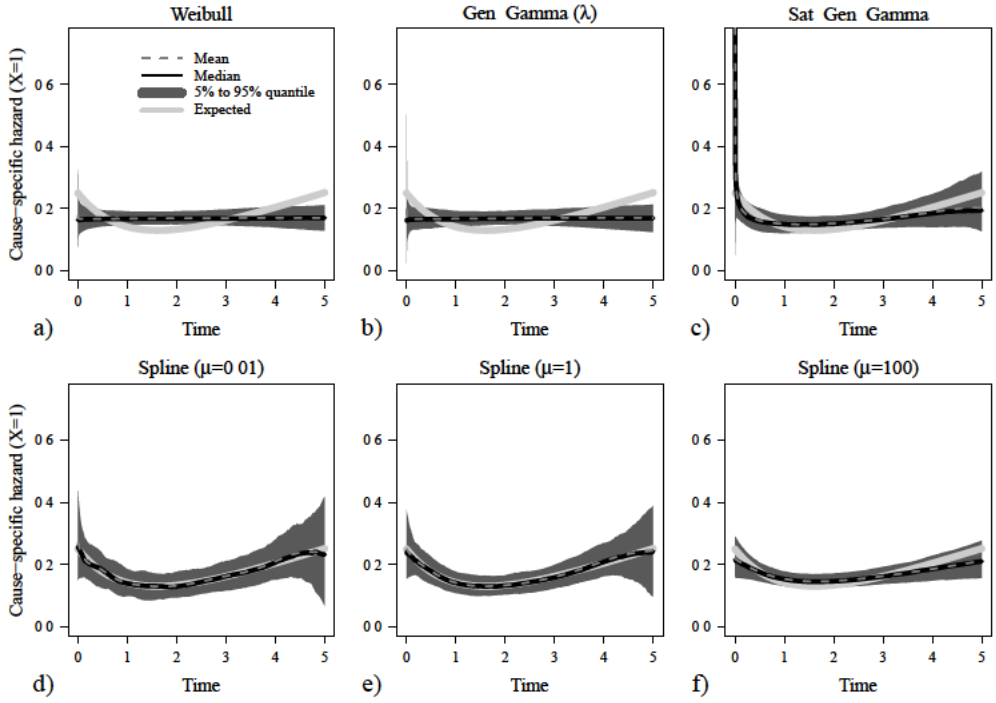

Figure B.32: Scenario III - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

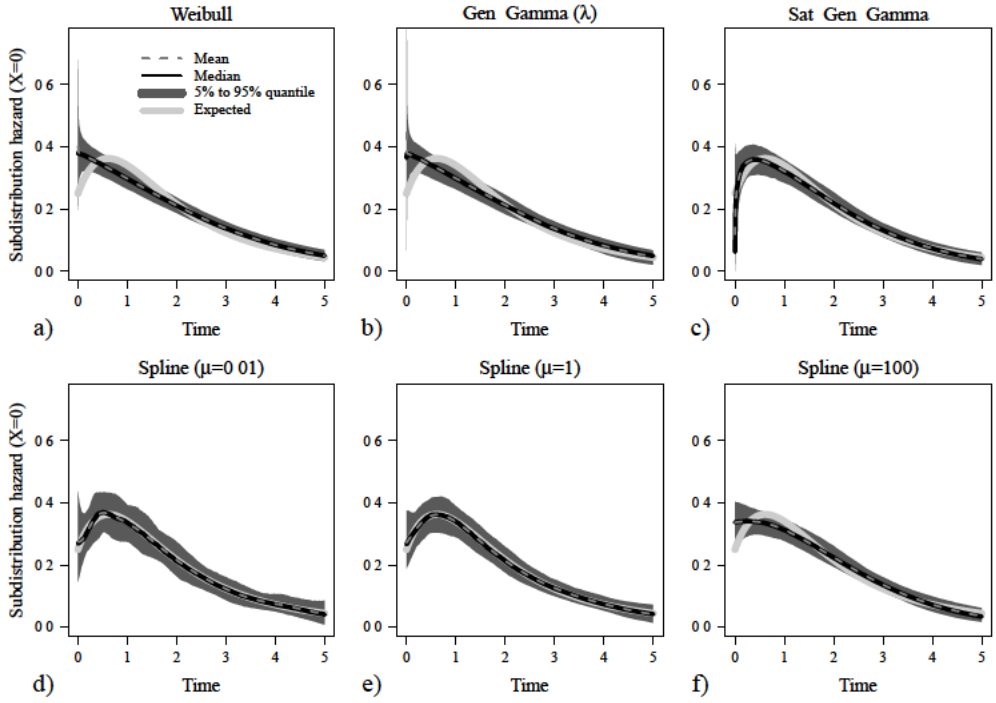

Figure B.33: Scenario III - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

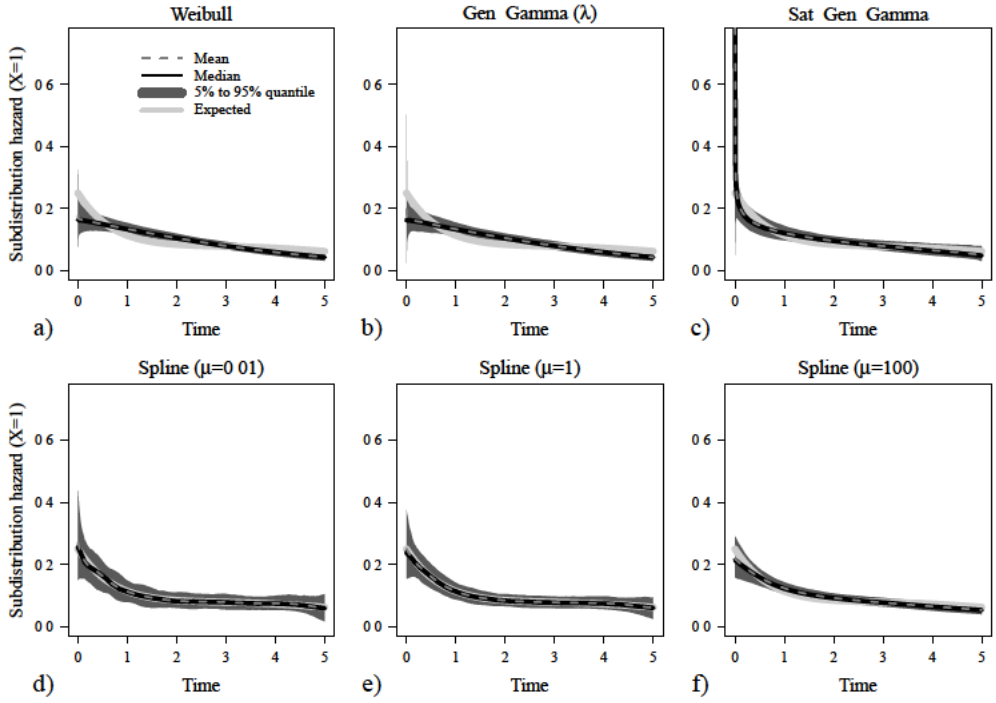

Figure B.34: Scenario III - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

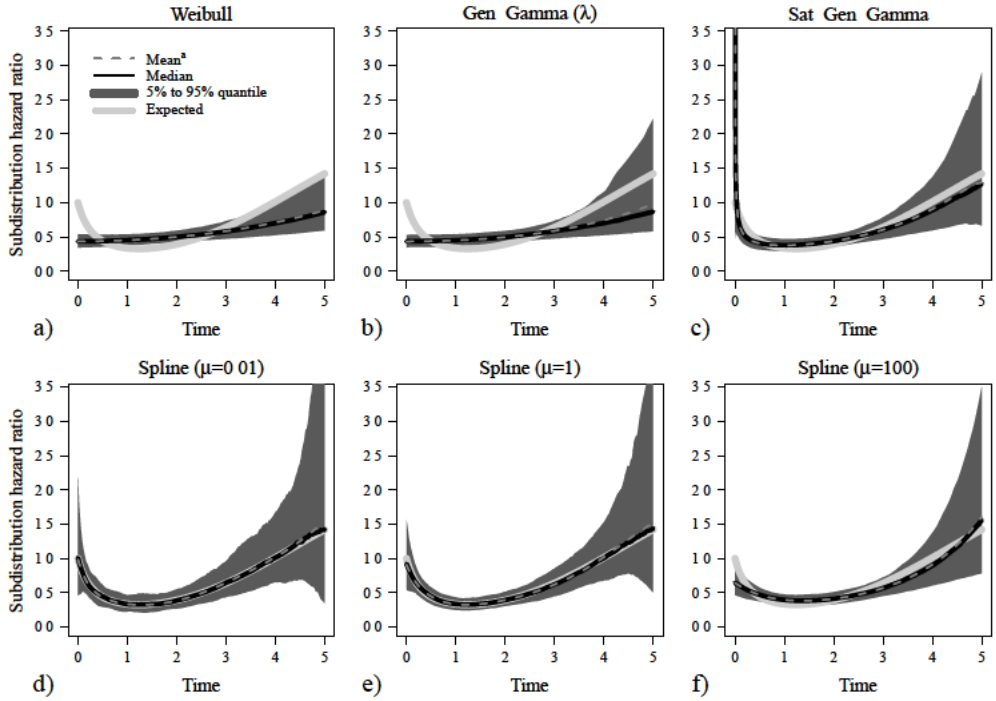

Figure B.35: Scenario III - low censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario III - Moderate Censoring

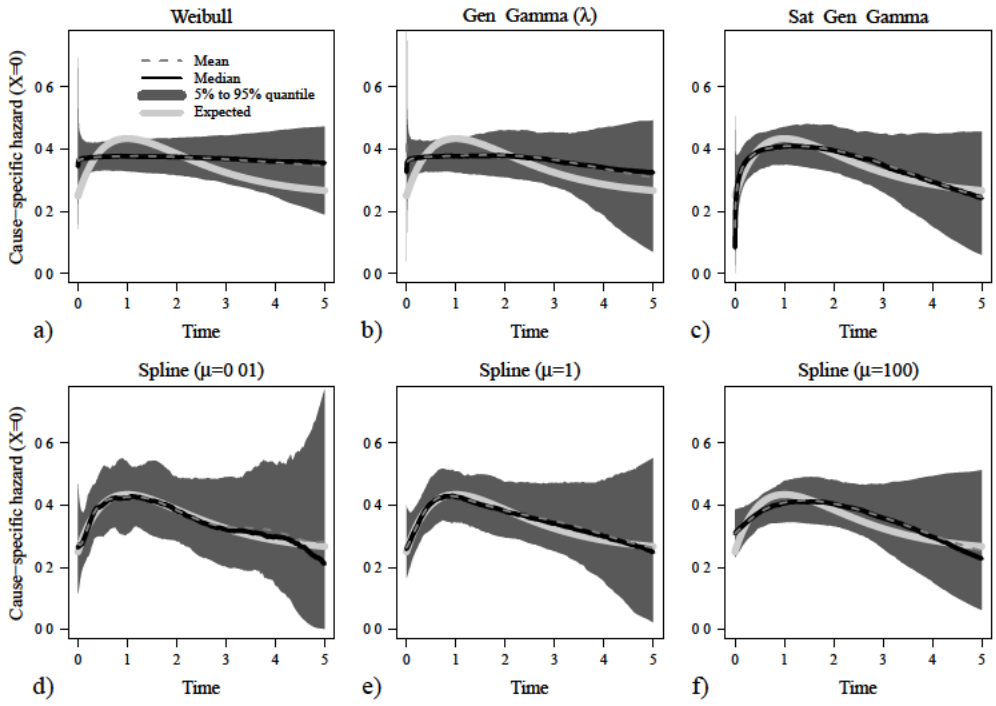

Figure B.36: Scenario III - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

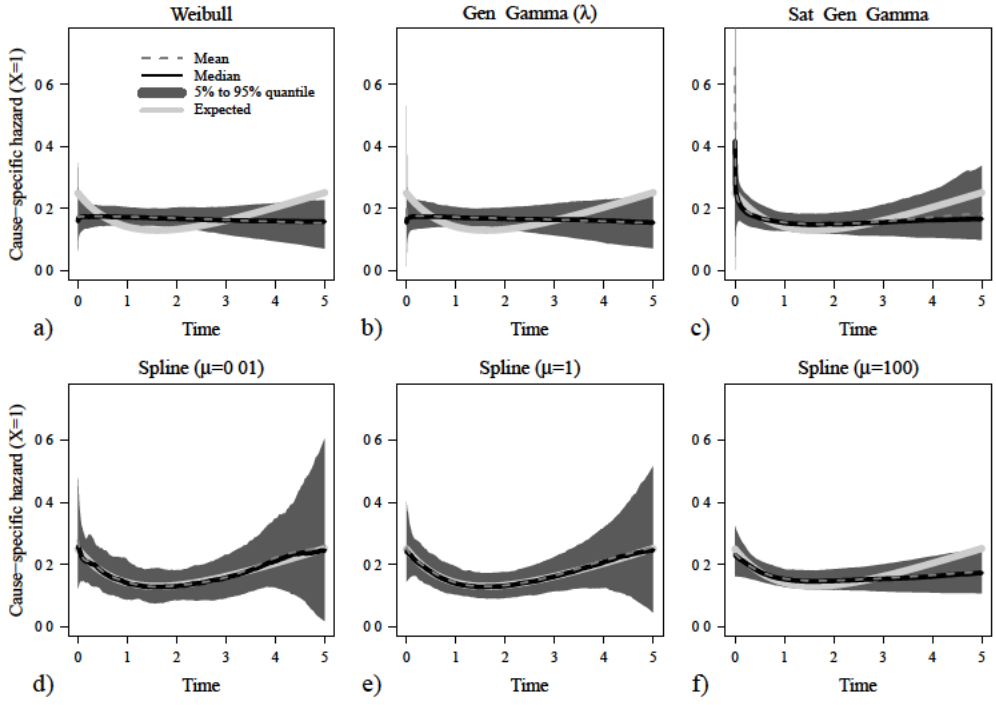

Figure B.37: Scenario III - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

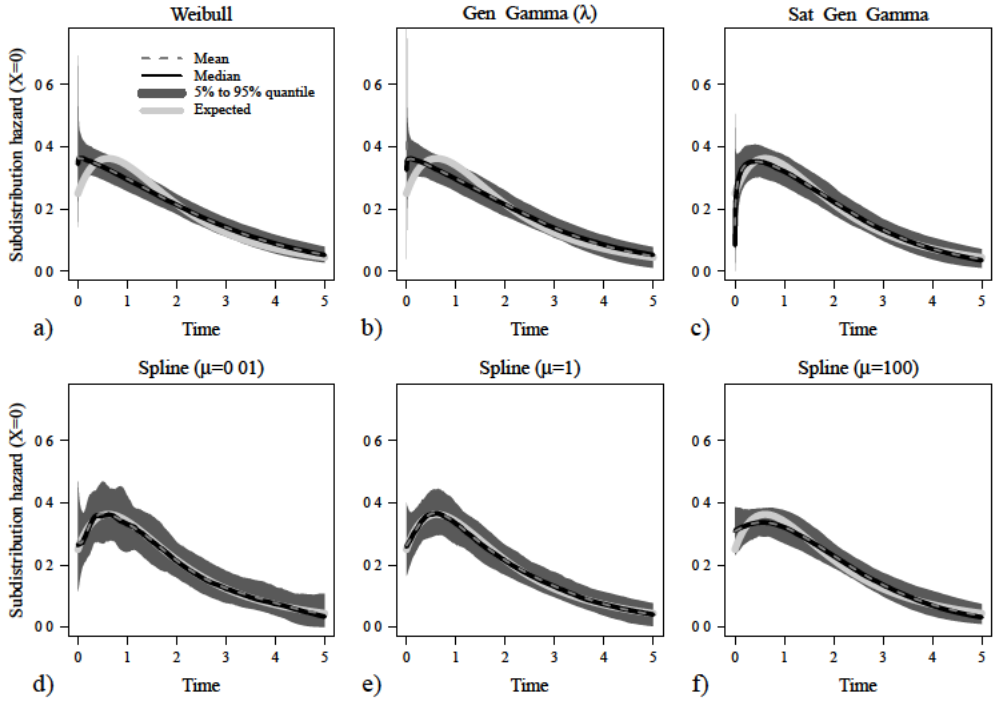

Figure B.38: Scenario III - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0).

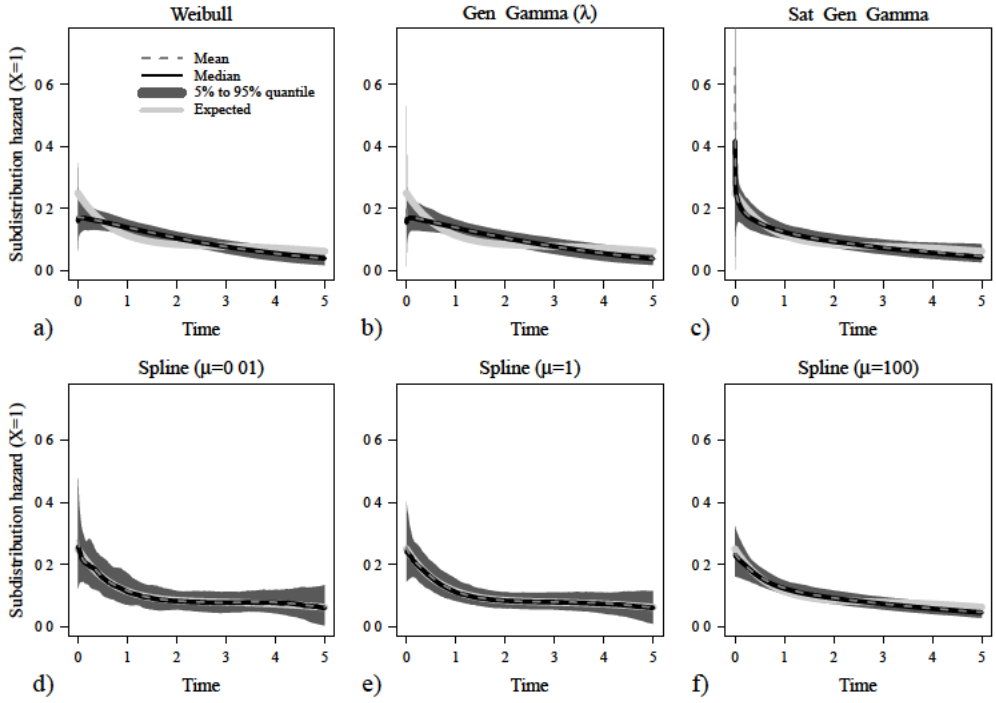

Figure B.39: Scenario III - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

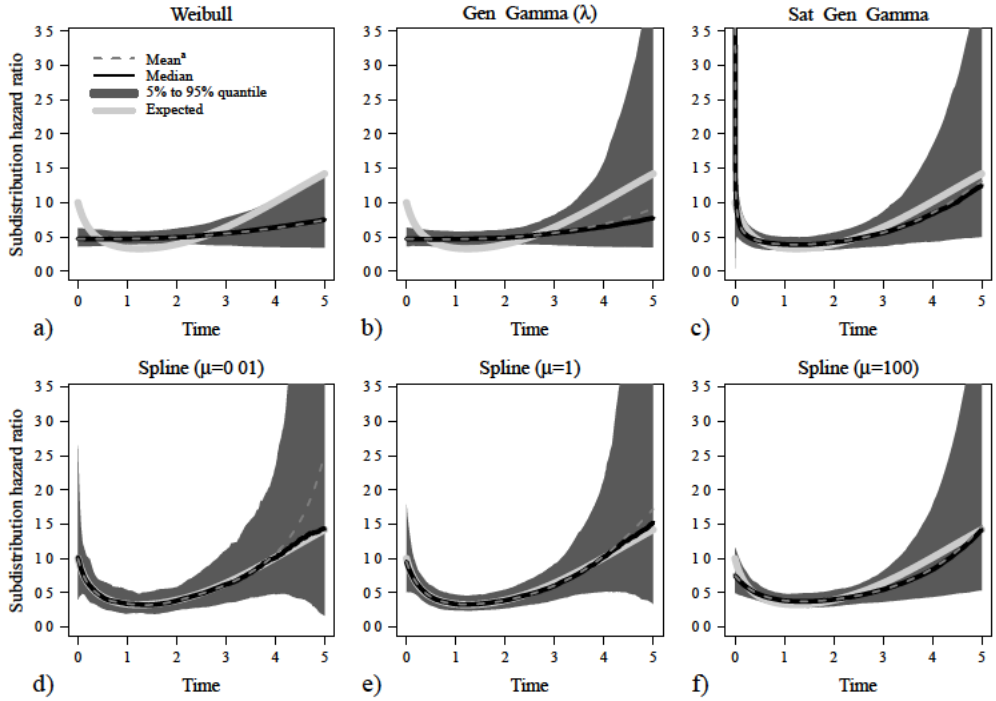

Figure B.40: Scenario III - moderate censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario III - High Censoring

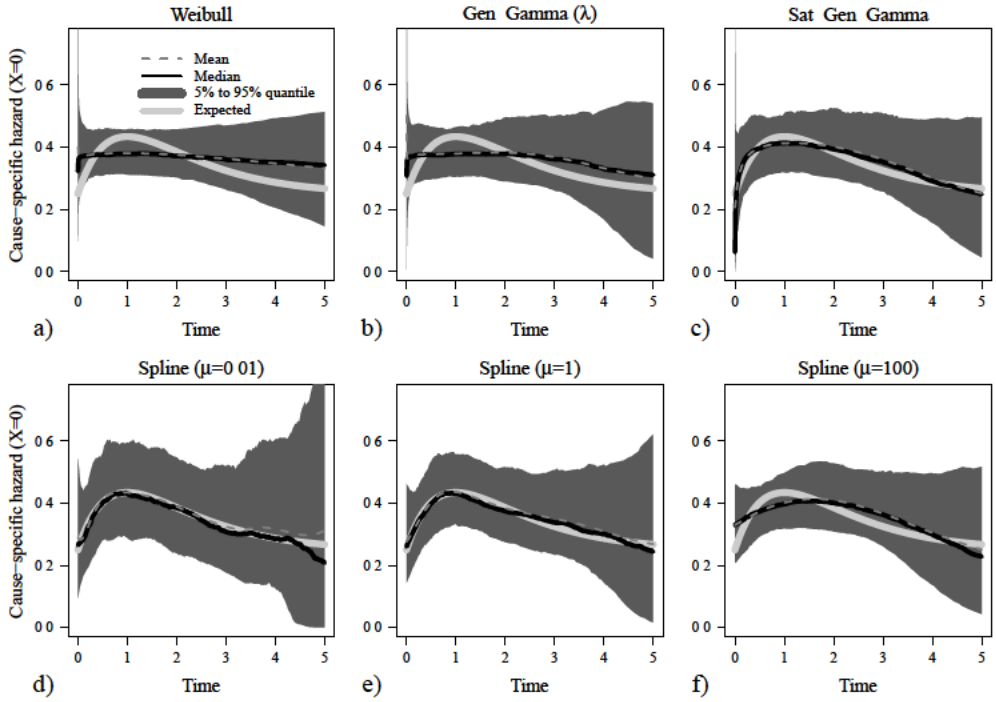

Figure B.41: Scenario III - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

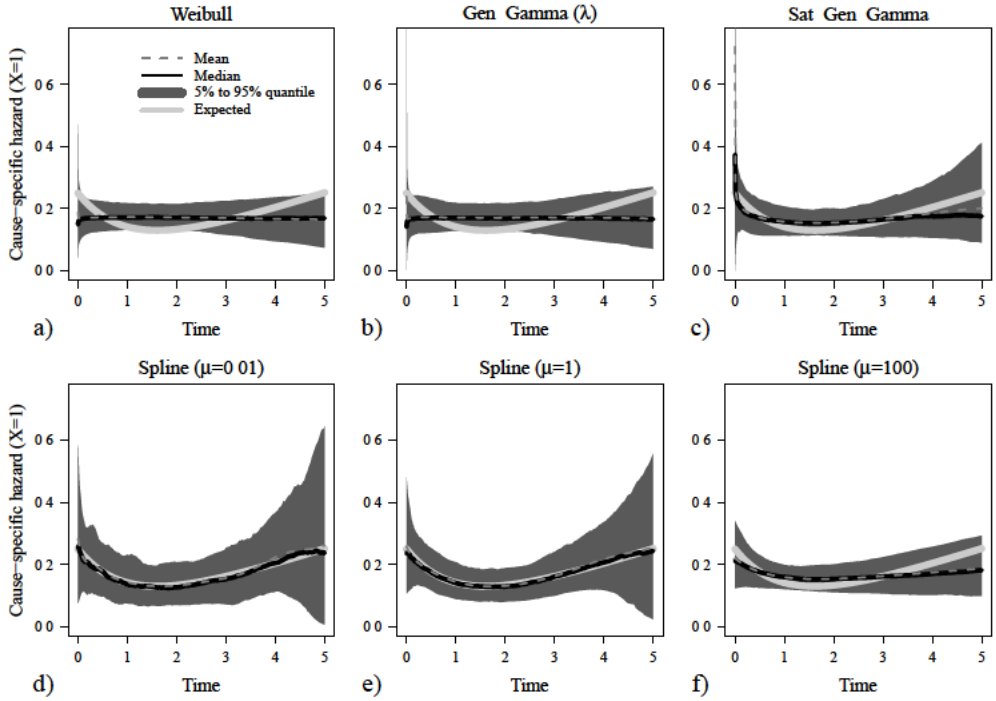

Figure B.42: Scenario III - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

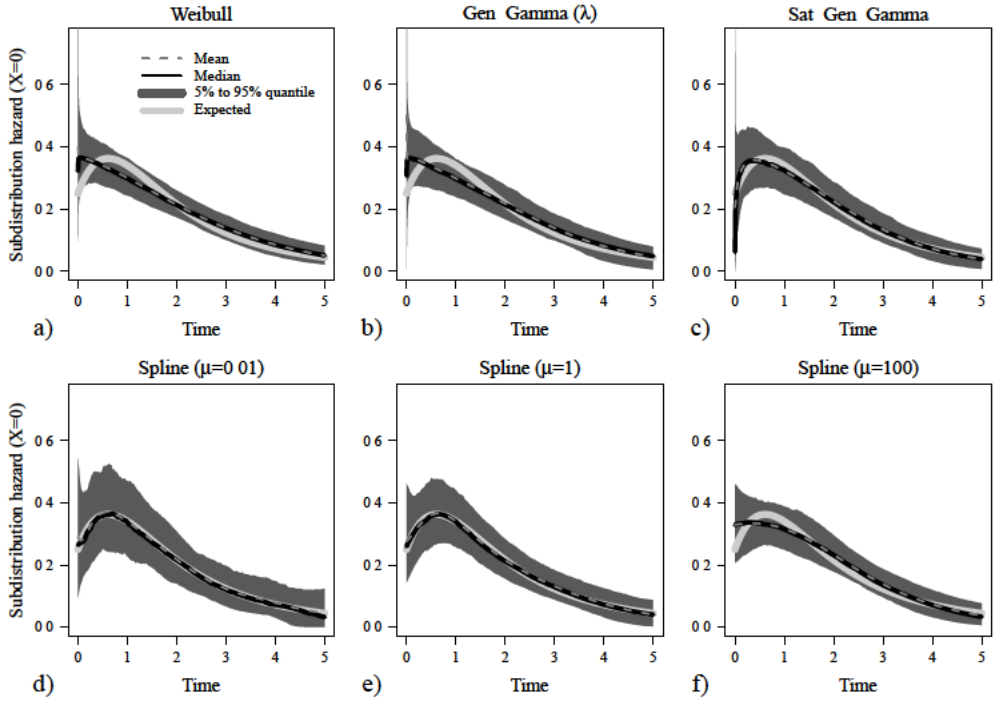

Figure B.43: Scenario III - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0)

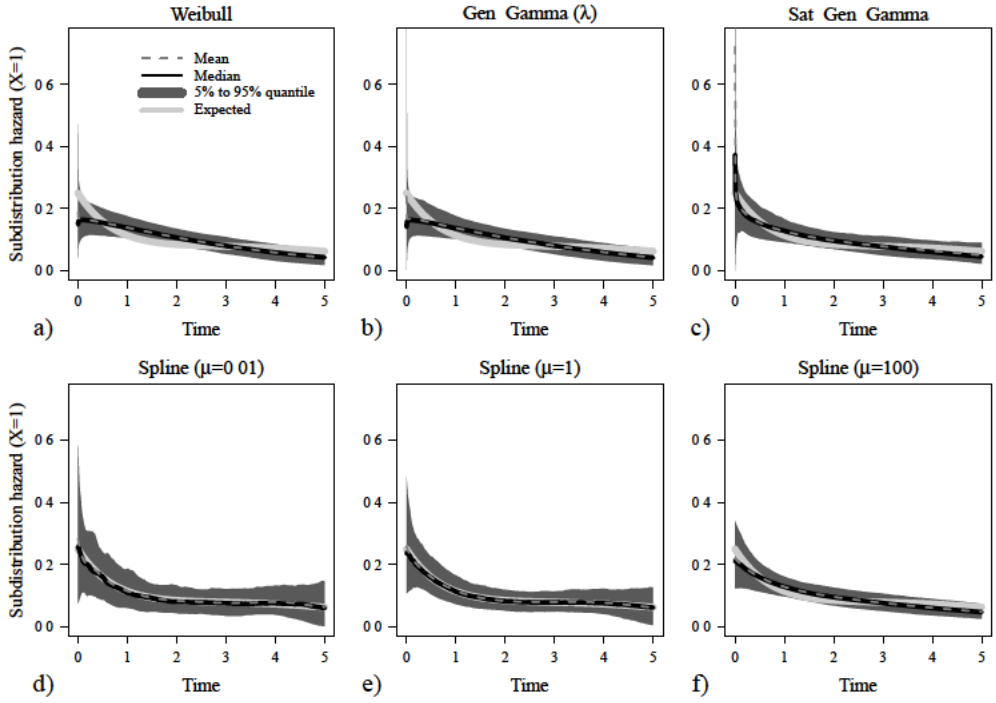

Figure B.44: Scenario III - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

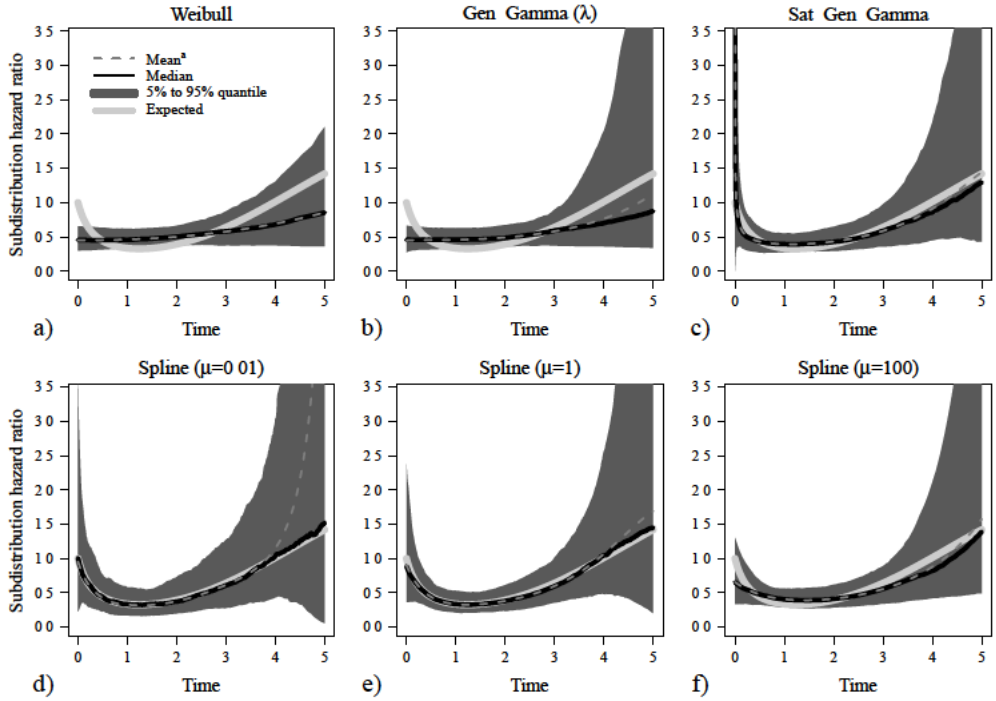

Figure B.45: Scenario III - high censoring: Estimated subdistribution hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

B.1.4 Constant subdistribution hazard ratio

Further results for the simulation scenario with a time-constant subdistribution hazard ratio, as described in Section 7.3.4 and Section 7.4.4, are shown here. The most important results, namely estimates for the average subdistribution (log-)hazard ratio and illustrations of the estimated subdistribution hazard ratios, are presented in Section 7.4.4. Summaries of estimates for the subdistribution hazards for the event of interest are shown here for both groups as well as summaries of estimates for the cause-specific hazards and hazard ratios.
Expected subdistribution and cause-specific hazard rates are illustrated in the according figures by solid grey lines. The expected cause-specific hazard ratio was derived as quotient of the expected cause-specific hazard rates.
Results using the different censoring distributions presented in Section 7.1 are displayed on the following pages:

- Low amount of censored observations - pages 156 to 158
- Moderate amount of censored observations - pages 159 to 161
- High amount of censored observations - pages 162 to 164

Scenario IV - Low Censoring

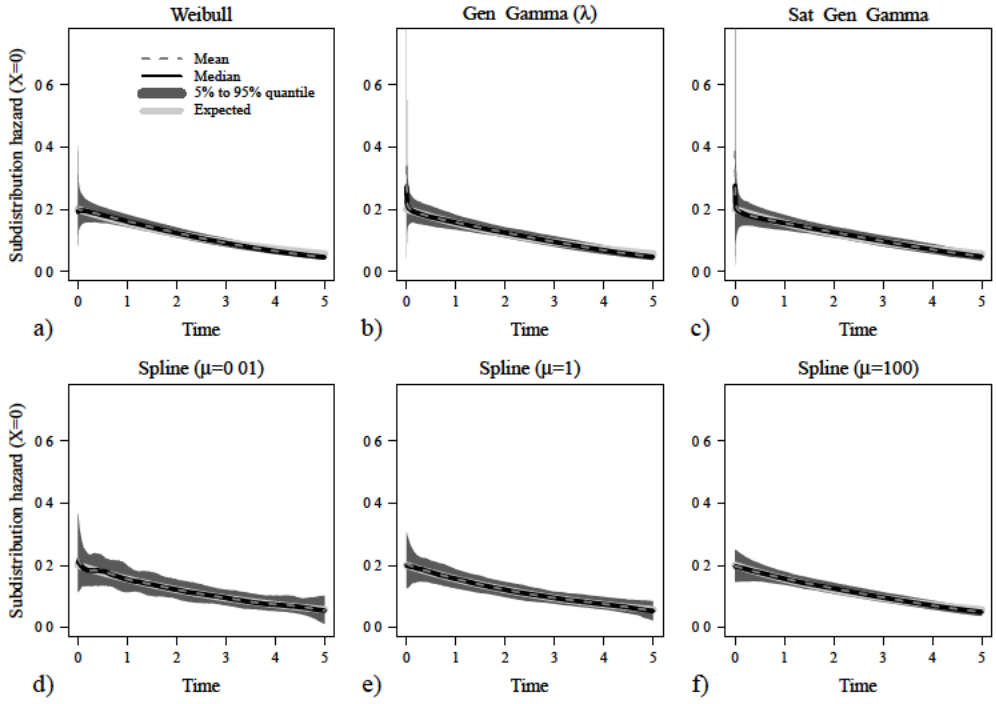

Figure B.46: Scenario IV - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0).

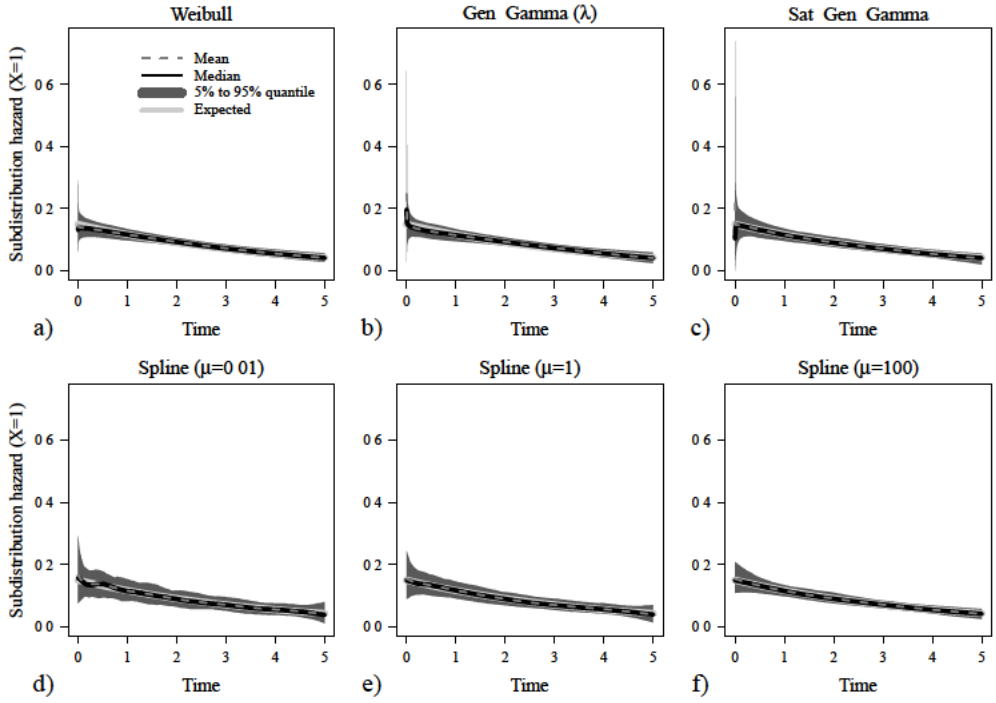

Figure B.47: Scenario IV - low censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

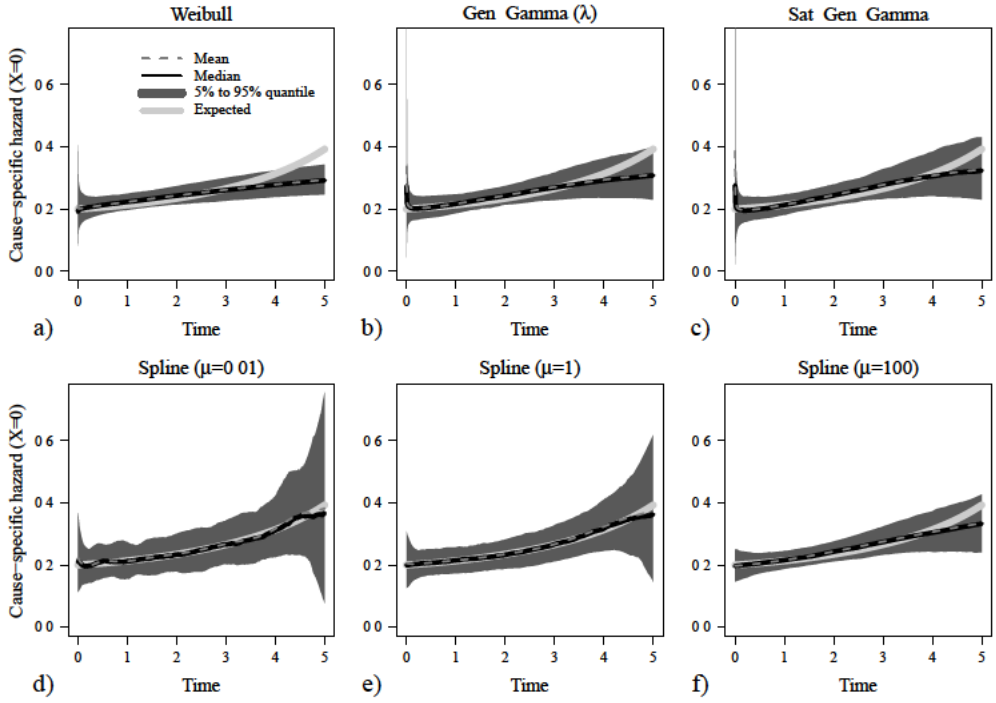

Figure B.48: Scenario IV - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

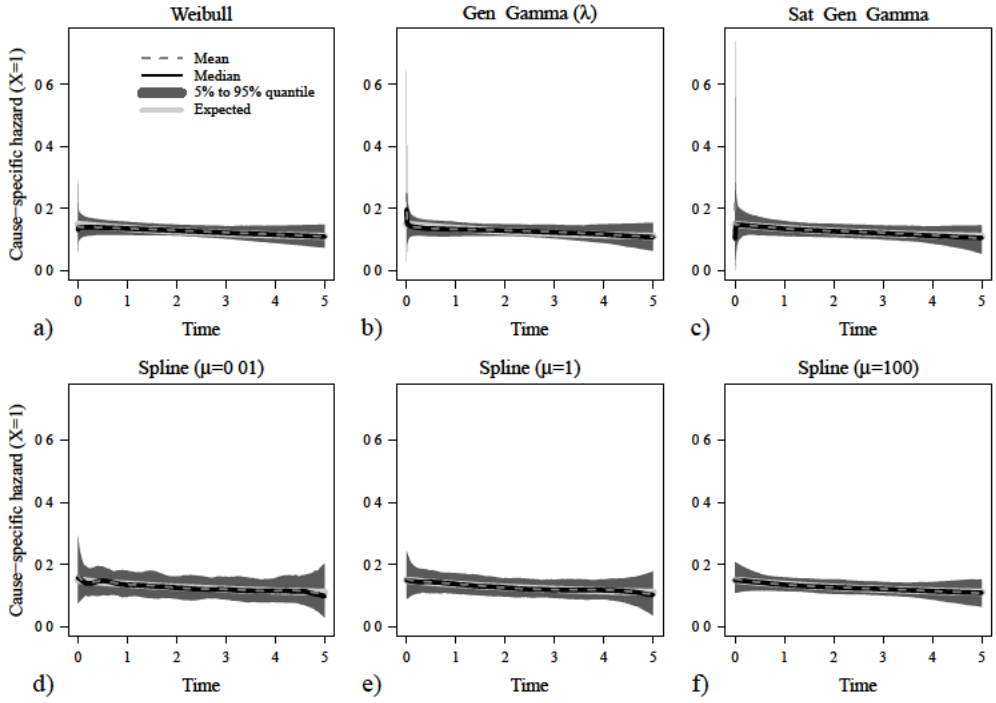

Figure B.49: Scenario IV - low censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

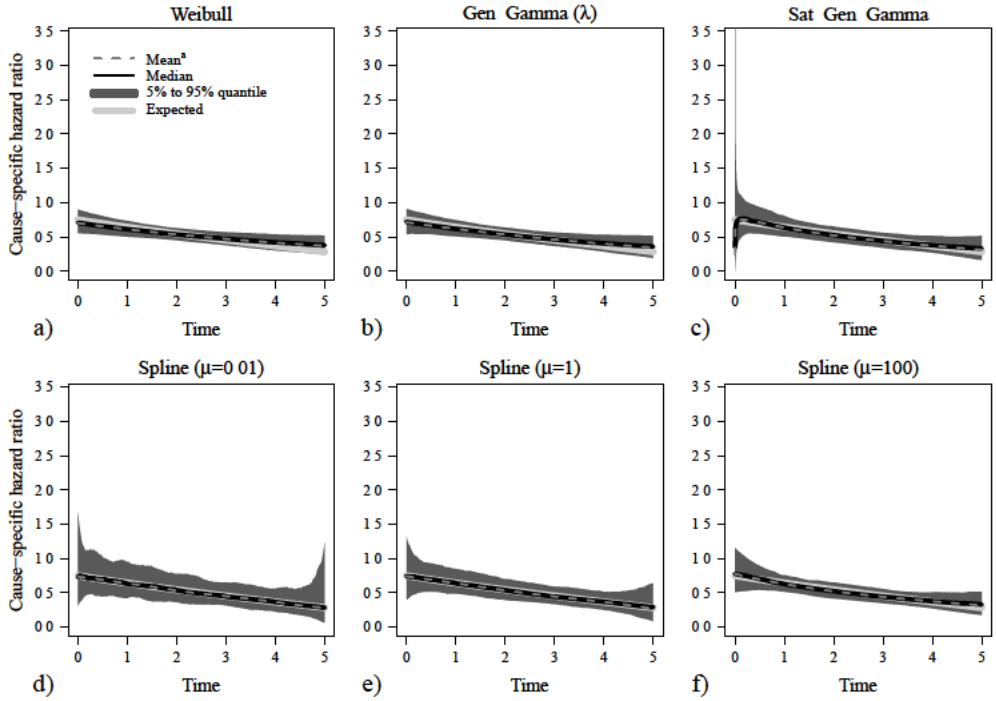

Figure B.50: Scenario IV - low censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario IV - Moderate Censoring

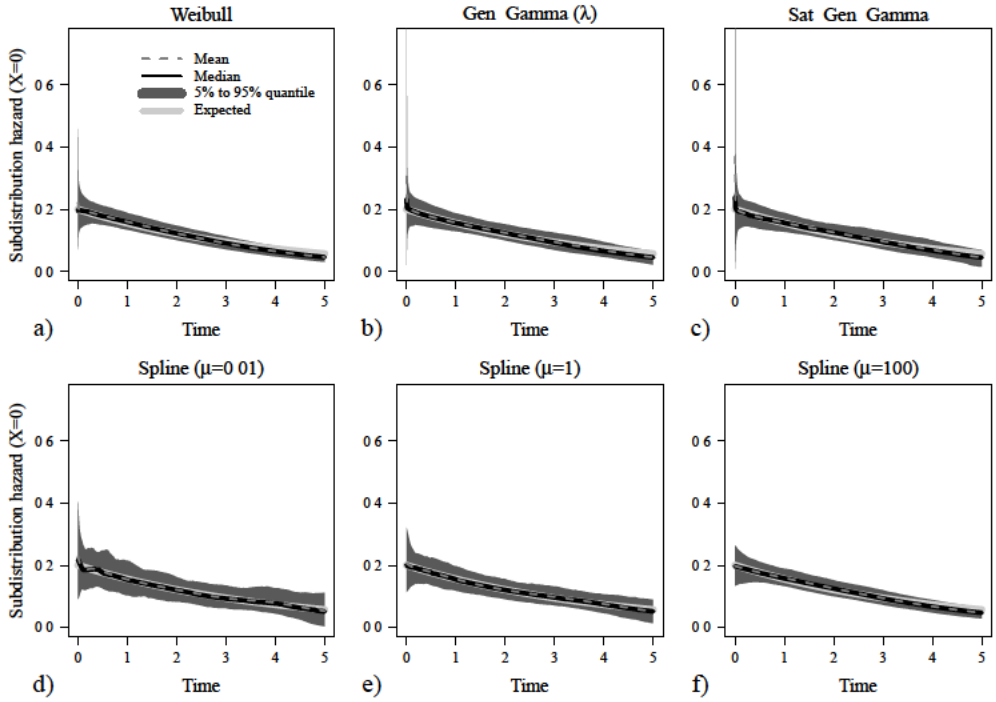

Figure B.51: Scenario IV - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0).

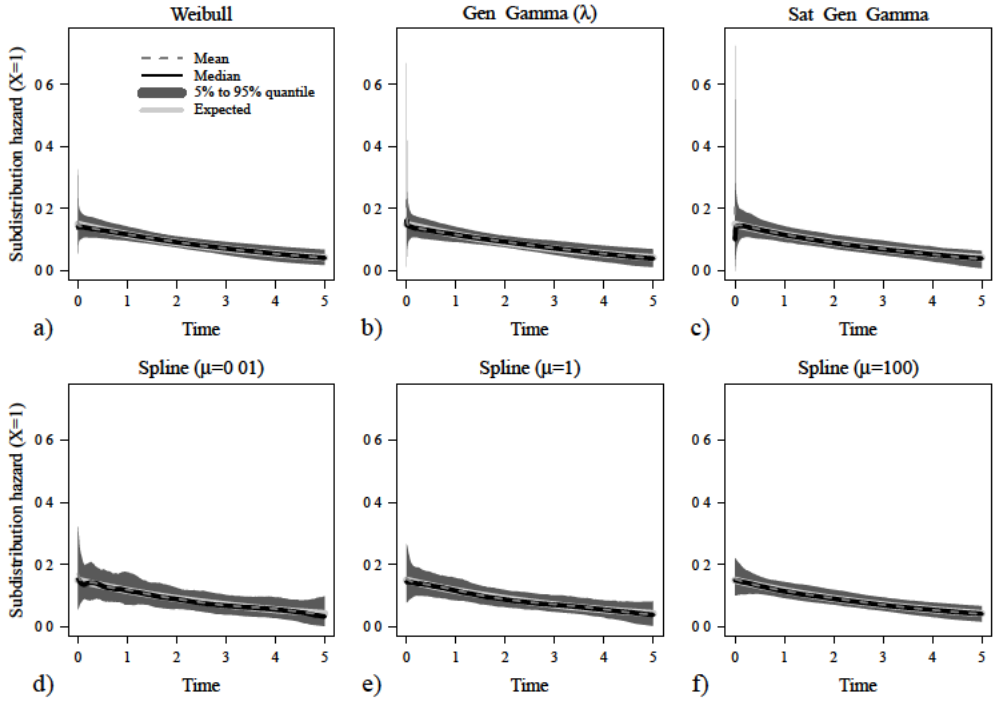

Figure B.52: Scenario IV - moderate censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

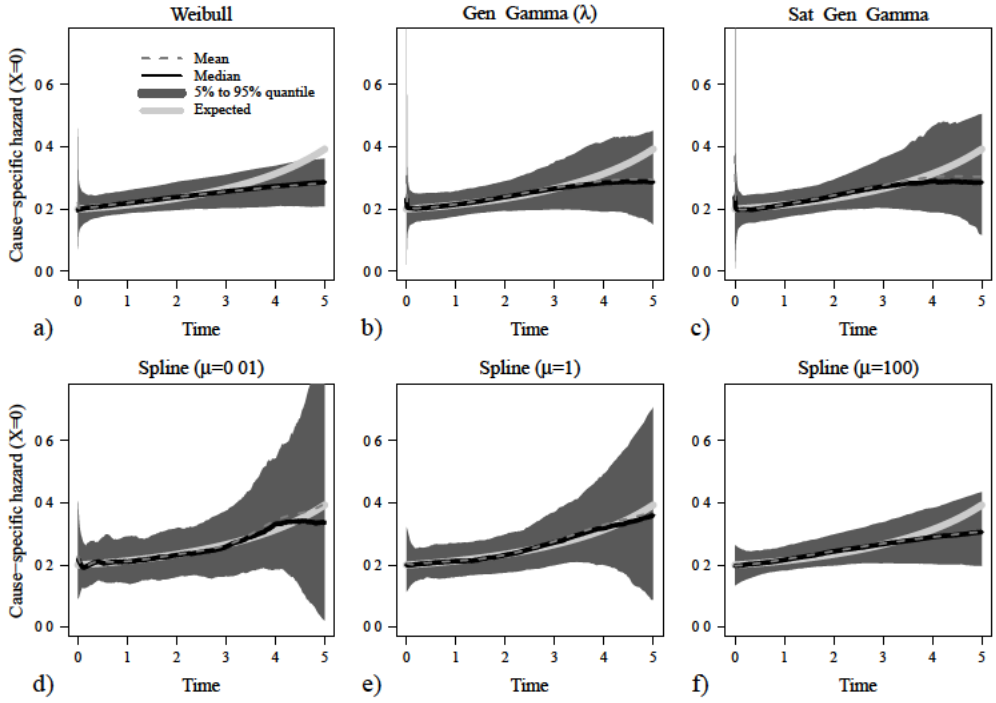

Figure B.53: Scenario IV - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

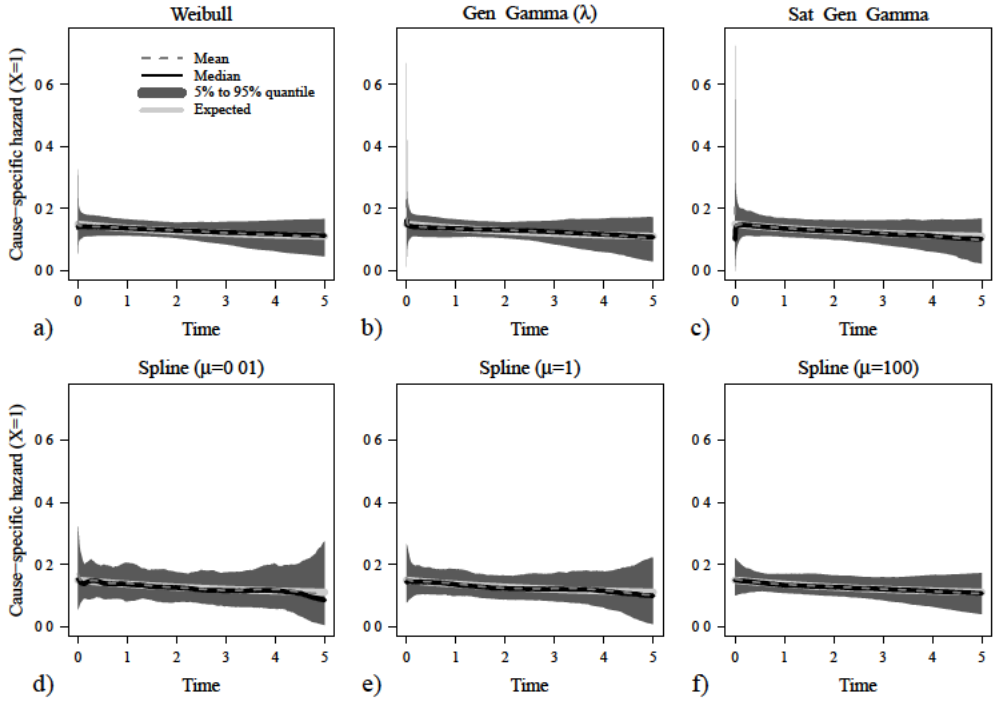

Figure B.54: Scenario IV - moderate censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

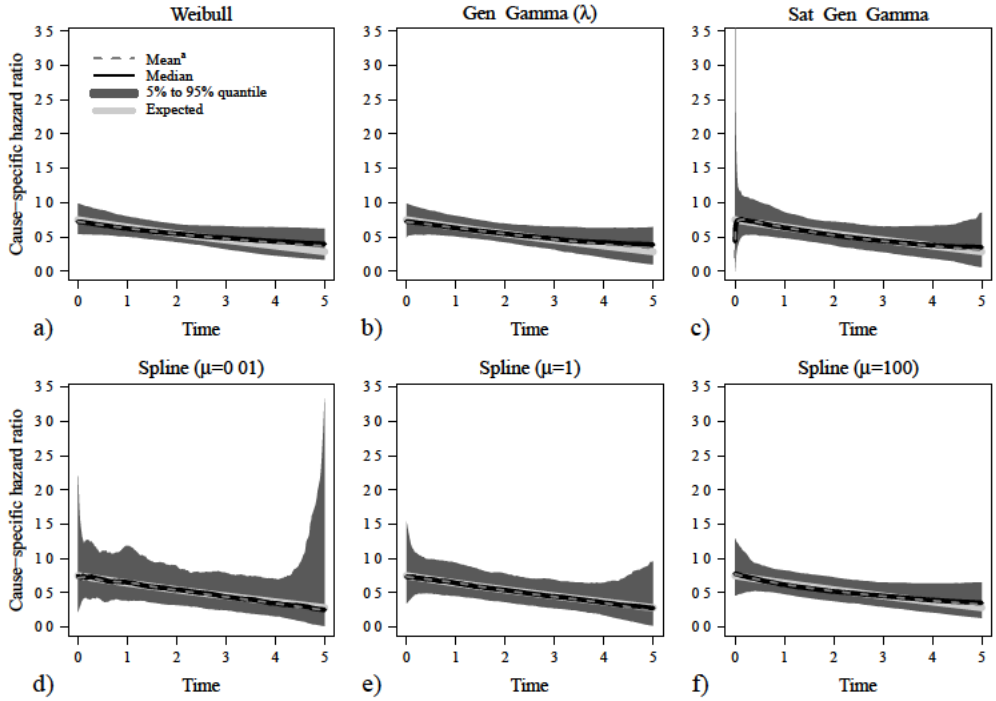

Figure B.55: Scenario IV - moderate censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

Scenario IV - High Censoring

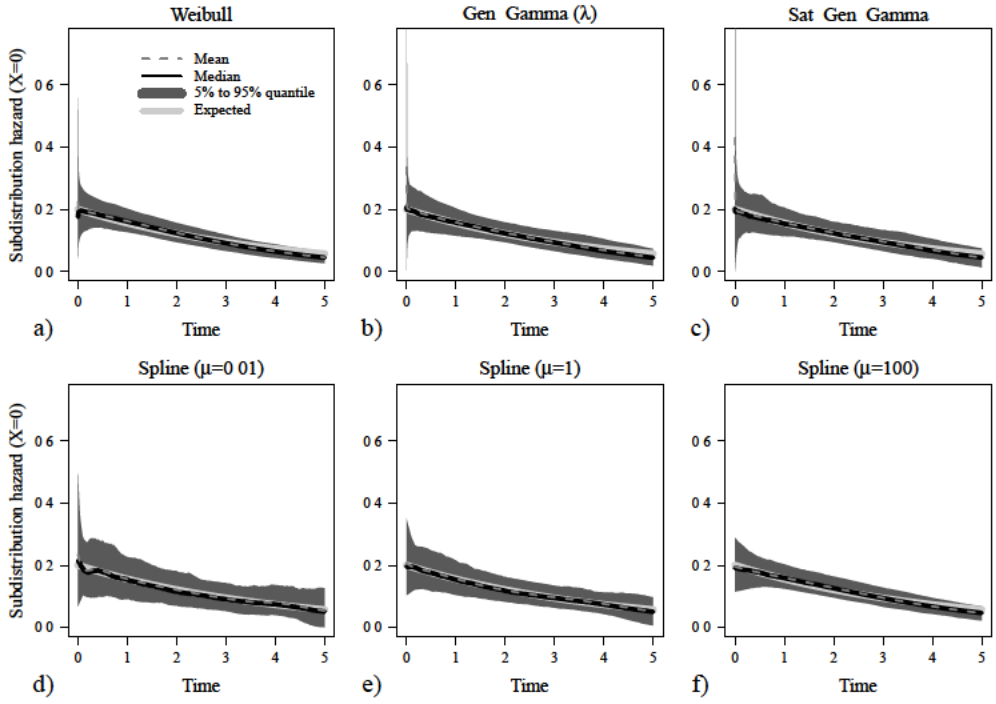

Figure B.56: Scenario IV - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the control group (X=0).

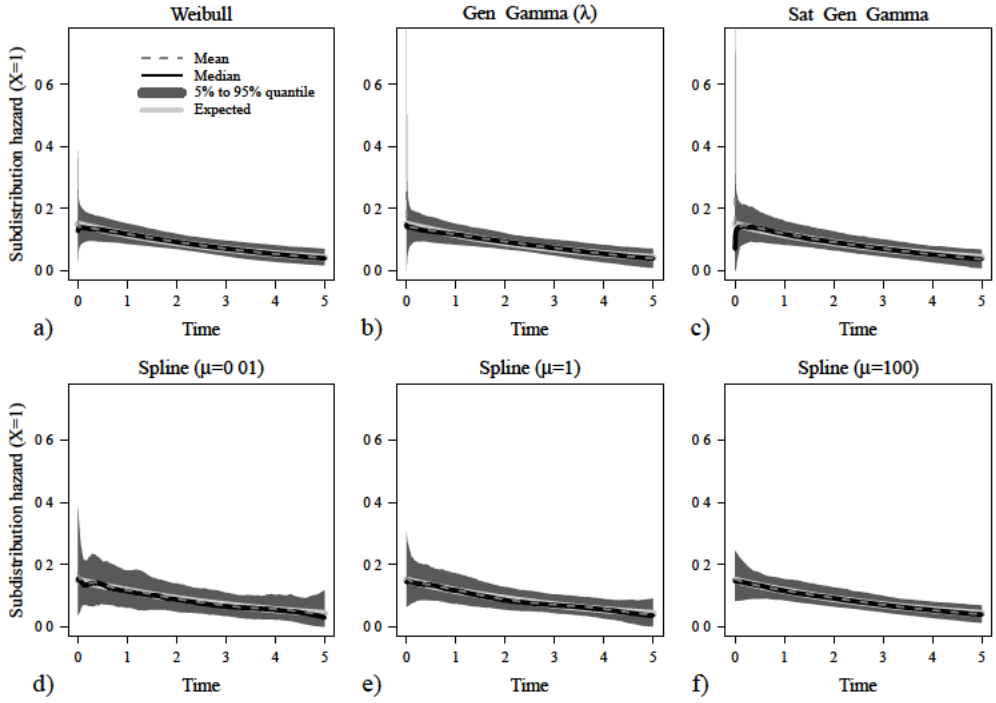

Figure B.57: Scenario IV - high censoring: Estimated subdistribution hazard rates for parametric and spline-based methods for the study group (X=1).

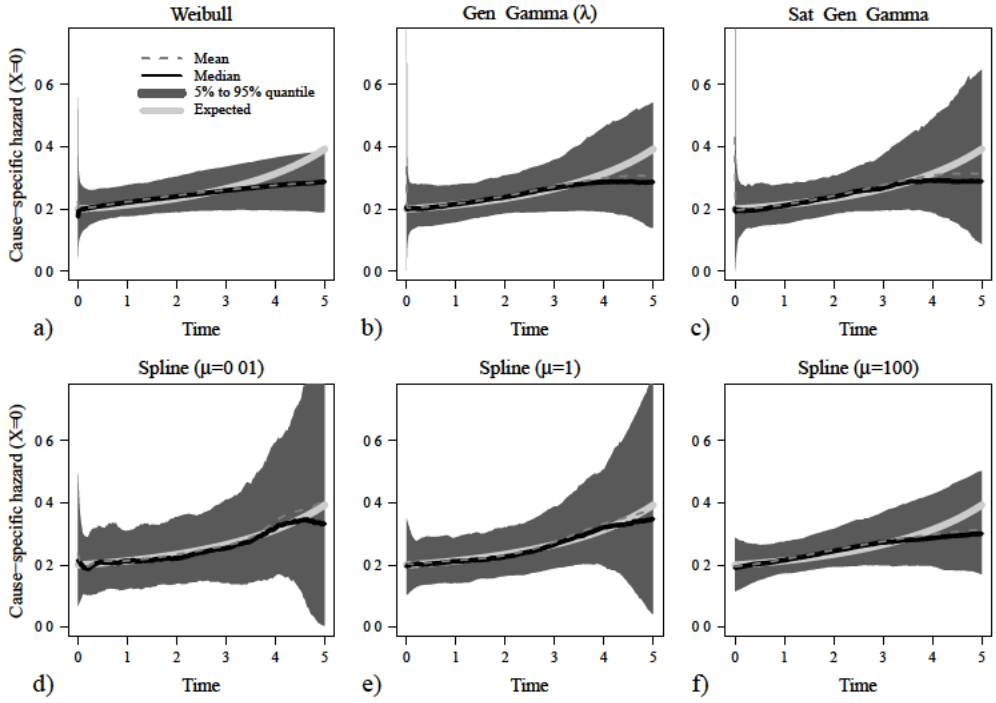

Figure B.58: Scenario IV - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the control group (X=0).

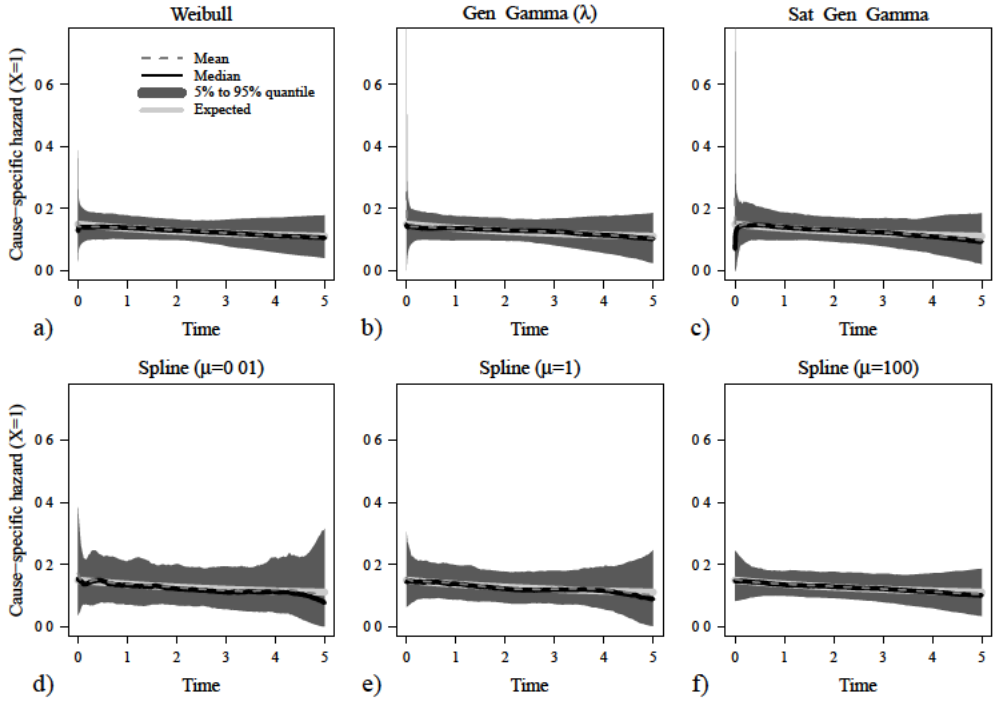

Figure B.59: Scenario IV - high censoring: Estimated cause-specific hazard rates for parametric and spline-based methods for the study group (X=1).

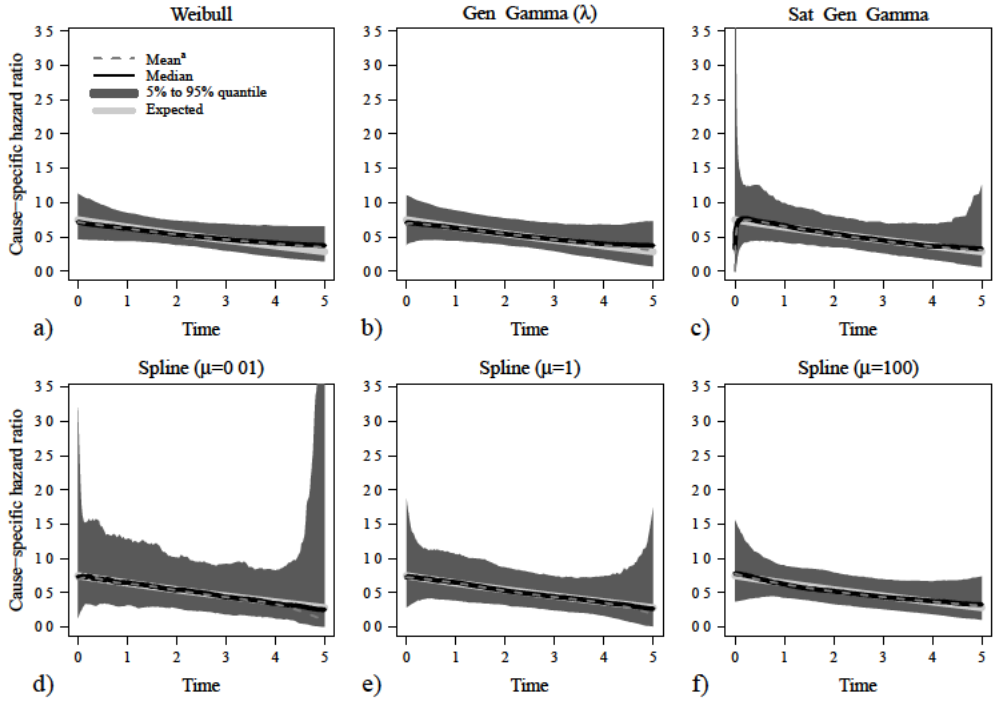

Figure B.60: Scenario IV - high censoring: Estimated cause-specific hazard ratios for parametric and spline-based methods.
[a]Mean hazard ratios were derived as exponentiated means of log-hazard ratio estimates.

B.2 Sketch of the R-code

B.2.1 Functions for simulation

Data generation

For illustration of the code used for data generation, the according function for Scenario III is presented. Variables $C1$ and $C2$ are parameters of the Weibull distribution for generation of potential censoring times and can be set as described in Section 7.1. It has to be considered, that the parametrization of the Weibull distribution used in *rweibull* is different from the parametrization used throughout this work (see Equation 2.21), and that $C1$ and $C2$ have to be adapted accordingly.

```
Gen.Data.Sz3 <- function(n,AdminCens,C1,C2)
{
group <- sample(0:1,n,repl=T)

# Definition of cause-specific hazard functions
# for the event of interest
h1_A <- function(zz) 0.25 + 0.5*zz / exp(zz)
h1_B <- function(zz) 0.25 / exp(zz) + 0.05*zz
# Cumulative cause-specific hazard functions
H1_A <- function(zz) 0.25 * exp(-zz) *((zz+2)*exp(zz)-2*zz-2)
H1_B <- function(zz) 0.025 * exp(-zz) * ((zz^2+10)*exp(zz)-10)

# Definition of cause-specific hazard functions
# for the competing event
h2_A <- 0.2
h2_B <- 0.2
H2_A <- function(zz) h2_A*zz
H2_B <- function(zz) h2_B*zz
S.fct_A <- function(z,U) exp(-H1_A(z)-H2_A(z)) - U
S.fct_B <- function(z,U) exp(-H1_B(z)-H2_B(z)) - U

# Generation of event times using the inversion method
ev.time <- c()
for(i in 1:n)
{
if(group[i]==0) {
 Uz <- runif(1)
 ev.time[i] <- uniroot(
  S.fct_A,c(0.00000000000001,500),U=Uz)$root}
if(group[i]==1){
 Uz <- runif(1)
 ev.time[i] <- uniroot(
  S.fct_B,c(0.00000000000001,500),U=Uz)$root}
}
```

```
# Determination of event types
ev.type <- c()
for(i in 1:n)
{
if(group[i]==0)
ev.type[i] <- sample(1:2,1,prob=c(h1_A(ev.time[i]),h2_A))
if(group[i]==1)
ev.type[i] <- sample(1:2,1,prob=c(h1_B(ev.time[i]),h2_B))
}

# Generation of censoring times
# and determination of status variable
 CT <- rweibull(n,C1,C2)
 CensTime <- pmin(rep(AdminCens,n),CT)
 obs.time <- pmin(CensTime,ev.time)
 stat <- ev.type * as.numeric(ev.time<CensTime)

# Sorting data
 ind <- sort(obs.time,index.return=T)$ix
 s.obs.time <- obs.time[ind]
 s.stat <- stat[ind]
 s.group <- group[ind]

# Generation of data frame
 Data <- data.frame(
  Time=s.obs.time,Status=s.stat,Group=s.group)
 return(Data)
}
```

Initiation of simulation runs

Code used for initiation of simulation runs calling the functions shown above and in Section B.2.2.

```
for(rr in 1:Runs)
{
# Generate and save data
if(Szen==1)
Data <- Gen.Data.Sz1(n=n,lam1A=1/3,lam1B=0.25,
   lam2A=0.2,lam2B=0.2,AdminCens=Admin.Time,C1=C1,C2=C2)
if(Szen==2)
Data <- Gen.Data.Sz2(n=n,AdminCens=5,C1=C1,C2=C2)
if(Szen==3)
Data <- Gen.Data.Sz3(n=n,AdminCens=5,C1=C1,C2=C2)
if(Szen==4)
Data <- Gen.Data.Sz4(n=n,AdminCens=5,C1=C1,C2=C2)
Data.List[[rr]] <- Data
```

```
##################################
# Estimate coefficients
##################################

# Exponential distribution
Estim.Expo <- EST.EXPO(Data,c(0,0,log(0.1),0,log(0.1),0))
Expo.Coef[rr,] <- Estim.Expo$estimate
Conv.Expo[rr] <- Estim.Expo$code
# Define starting values for Weibull distribution
SV.Weibull <- c(Estim.Expo$est[1],Estim.Expo$est[2],
      Estim.Expo$est[3],Estim.Expo$est[4],1,
      Estim.Expo$est[5],Estim.Expo$est[6],1)

# Weibull distribution
Estim.Weibull <- EST.WEIBULL(Data,SV.Weibull)
Weibull.Coef[rr,] <- Estim.Weibull$estimate
Conv.Weibull[rr] <- Estim.Weibull$code
# Define starting values for gen. gamma distribution
SV.Ggv <- c(Estim.Weibull$est[1],Estim.Weibull$est[2],
      Estim.Weibull$est[3],Estim.Weibull$est[4],
      1/Estim.Weibull$est[5],1,
      Estim.Weibull$est[6],Estim.Weibull$est[7],
      1/Estim.Weibull$est[8],1)

# Gen. gamma distribution
Estim.Ggv <- EST.GGV(Data,SV.Ggv)
Ggv.Coef[rr,] <- Estim.Ggv$estimate
Conv.Ggv[rr] <- Estim.Ggv$code
if(Estim.Ggv$code==1)
# Define starting values for sat. gen. gamma distribution
SV.Sat.Ggv <- c(Estim.Ggv$est[1],Estim.Ggv$est[2],
        Estim.Ggv$est[3],Estim.Ggv$est[4],
Estim.Ggv$est[5],0,Estim.Ggv$est[6],0,
        Estim.Ggv$est[7],Estim.Ggv$est[8],
        Estim.Ggv$est[9],0,Estim.Ggv$est[10],0)
if(Estim.Ggv$code!=1)
SV.Sat.Ggv <- c(Estim.Weibull$est[1],Estim.Weibull$est[2],
        Estim.Weibull$est[3],Estim.Weibull$est[4],
1/Estim.Weibull$est[5],0,1,0,
        Estim.Weibull$est[6],Estim.Weibull$est[7],
        1/Estim.Weibull$est[8],0,1,0)

# Sat. gen. gamma distribution
Estim.Sat.Ggv <- EST.SAT.GGV(Data,SV.Sat.Ggv)
Sat.Ggv.Coef[rr,] <- Estim.Sat.Ggv$estimate
Conv.Sat.Ggv[rr] <- Estim.Sat.Ggv$code
```

```
# Define starting values for spline approach
SV.Spline <- c(Estim.Expo$est[1:2],
       rep(-0.5,N.InnerKnots+4),rep(0,N.InnerKnots+4),
       rep(-0.5,N.InnerKnots+4),rep(0,N.InnerKnots+4))

# Spline approach (mu=0.01)
Estim.Splines_0.01 <- EST.SPLINE(Data,mu=0.01,
       N.innerKnots=N.InnerKnots,Start.Vals=SV.Spline)
Splines.Coef_0.01[rr,] <- Estim.Splines_0.01$estimate
Conv.Splines_0.01[rr] <- Estim.Splines_0.01$code

# Spline approach (mu=1)
Estim.Splines_1 <- EST.SPLINE(Data,mu=1,
       N.innerKnots=N.InnerKnots,Start.Vals=SV.Spline)
Splines.Coef_1[rr,] <- Estim.Splines_1$estimate
Conv.Splines_1[rr] <- Estim.Splines_1$code

# Spline approach (mu=100)
Estim.Splines_100 <- EST.SPLINE(Data,mu=100,
       N.innerKnots=N.InnerKnots,Start.Vals=SV.Spline)
Splines.Coef_100[rr,] <- Estim.Splines_100$estimate
Conv.Splines_100[rr] <- Estim.Splines_100$code


#########################################################
# Derive cause-specific and
# subdistribution hazard rates for observed times
#########################################################

Haz.Weibull.ObsTime[[rr]] <- Deriv.Weibull(
     Timepoints=Data.List[[rr]]$Time,theta=Weibull.Coef[rr,])
Haz.Ggv.ObsTime[[rr]] <- Deriv.Ggv(
     Timepoints=Data.List[[rr]]$Time,theta=Ggv.Coef[rr,])
Haz.Sat.Ggv.ObsTime[[rr]] <- Deriv.Sat.Ggv(
     Timepoints=Data.List[[rr]]$Time,theta=Sat.Ggv.Coef[rr,])
Haz.Splines_0.01.ObsTime[[rr]] <- Deriv.Splines(
     Basis=Bsplines.quant(TimeEval=Data$Time,
     TimeKnots=Data$Time,Status=Data$Status,IB=TRUE,
     n.Knots=N.InnerKnots+2),theta=Splines.Coef_0.01[rr,])
Haz.Splines_1.ObsTime[[rr]] <- Deriv.Splines(
     Basis=Bsplines.quant(TimeEval=Data$Time,
     TimeKnots=Data$Time,Status=Data$Status,IB=TRUE,
     n.Knots=N.InnerKnots+2),theta=Splines.Coef_1[rr,])
Haz.Splines_100.ObsTime[[rr]] <- Deriv.Splines(
     Basis=Bsplines.quant(TimeEval=Data$Time,
     TimeKnots=Data$Time,Status=Data$Status,IB=TRUE,
     n.Knots=N.InnerKnots+2),theta=Splines.Coef_100[rr,])
```

```
#########################################################
# Derive cause-specific and
# subdistribution hazard rates for given timepoints
#########################################################

Haz.Expo.GivenTime[[rr]] <- Deriv.Expo(
       Timepoints=Given.Times, theta=Expo.Coef[rr,])
Haz.Weibull.GivenTime[[rr]] <- Deriv.Weibull(
       Timepoints=Given.Times, theta=Weibull.Coef[rr,])
Haz.Ggv.GivenTime[[rr]] <- Deriv.Ggv(
       Timepoints=Given.Times, theta=Ggv.Coef[rr,])
Haz.Sat.Ggv.GivenTime[[rr]] <- Deriv.Sat.Ggv(
       Timepoints=Given.Times, theta=Sat.Ggv.Coef[rr,])
Haz.Splines_0.01.GivenTime[[rr]] <- Deriv.Splines(
       Basis=Bsplines.quant(TimeEval=Given.Times,
       TimeKnots=Data$Time,Status=Data$Status,IB=TRUE,
       n.Knots=N.InnerKnots+2),theta=Splines.Coef_0.01[rr,])
Haz.Splines_1.GivenTime[[rr]] <- Deriv.Splines(
       Basis=Bsplines.quant(TimeEval=Given.Times,
       TimeKnots=Data$Time,Status=Data$Status,IB=TRUE,
       n.Knots=N.InnerKnots+2),theta=Splines.Coef_1[rr,])
Haz.Splines_100.GivenTime[[rr]] <- Deriv.Splines(
       Basis=Bsplines.quant(TimeEval=Given.Times,
       TimeKnots=Data$Time,Status=Data$Status,IB=TRUE,
       n.Knots=N.InnerKnots+2),theta=Splines.Coef_100[rr,])

print(rr)
}
```

B.2.2 Analysis of simulation runs

Log-likelihood functions

Code for the log-likelihood functions, which are to be maximized for estimation of the regression coefficients.

```
# Exponential mixture model
Loglik.Expo <- function(theta,Time,Status,Group){
  bpi0  <- theta[1]
  bpix  <- theta[2]
  b0_1  <- theta[3]
  bb_1  <- theta[4]
  b0_2  <- theta[5]
  bb_2  <- theta[6]

  LoLi <- sum(
  as.numeric(Status==1)*(
```

```
   ln.Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group) +
   Expo.log.dens(b0=b0_1,b1=bb_1,X=Group,ti=Time)) +
  as.numeric(Status==2)*(
   ln.Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group) +
   Expo.log.dens(b0=b0_2,b1=bb_2,X=Group,ti=Time)) +
  as.numeric(Status==0)*
   log(Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group)*
   Expo.Surv(b0=b0_1,bb_1,X=Group,ti=Time) +
        Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group)*
        Expo.Surv(b0=b0_2,bb_2,X=Group,ti=Time))
  )
  return(-LoLi)
}

# Weibull mixture model
Loglik.Weibull <- function(theta,Time,Status,Group){
  bpi0  <- theta[1]
  bpix  <- theta[2]
  b0_1  <- theta[3]
  bb_1  <- theta[4]
  alpha_1 <- theta[5]
  b0_2  <- theta[6]
  bb_2  <- theta[7]
  alpha_2 <- theta[8]

  LoLi <- sum(
  as.numeric(Status==1)*(
   ln.Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group) +
   Weibull.log.dens(b0=b0_1,b1=bb_1,
   alpha=alpha_1,X=Group,ti=Time)) +
  as.numeric(Status==2)*(
   ln.Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group) +
   Weibull.log.dens(b0=b0_2,b1=bb_2,
   alpha=alpha_2,X=Group,ti=Time)) +
  as.numeric(Status==0)*log(
   Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group)*
   Weibull.Surv(b0=b0_1,bb_1,
   alpha=alpha_1,X=Group,ti=Time) +
        Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group)*
        Weibull.Surv(b0=b0_2,bb_2,
         alpha=alpha_2,X=Group,ti=Time))
  )
  return(-LoLi)
}

# Generalized gamma (lambda) mixture model
Loglik.Ggv <- function(theta,Time,Status,Group){
  bpi0  <- theta[1]
  bpix  <- theta[2]
```

```
  b0_1  <- theta[3]
  bb_1  <- theta[4]
  a.tilde_1 <- theta[5]
  nu_1 <- theta[6]
  b0_2  <- theta[7]
  bb_2  <- theta[8]
  a.tilde_2 <- theta[9]
  nu_2 <- theta[10]

  LoLi <- sum(
  as.numeric(Status==1)*(
   ln.Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group) +
   Ggv.log.dens(b0=b0_1,b1=bb_1,a.tilde=a.tilde_1,
   nu=nu_1,X=Group,ti=Time)) +
  as.numeric(Status==2)*(
   ln.Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group) +
   Ggv.log.dens(b0=b0_2,b1=bb_2,a.tilde=a.tilde_2,
   nu=nu_2,X=Group,ti=Time)) +
  as.numeric(Status==0)*log(
        Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group)*
   Ggv.Surv(b0=b0_1,b1=bb_1,a.tilde=a.tilde_1,
   nu=nu_1,X=Group,ti=Time) +
        Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group)*
        Ggv.Surv(b0=b0_2,b1=bb_2,a.tilde=a.tilde_2,
         nu=nu_2,X=Group,ti=Time))
  )

  return(-LoLi)
}

# Saturated generalized gamma mixture model
Loglik.Sat.Ggv <- function(theta,Time,Status,Group){
  bpi0  <- theta[1]
  bpix  <- theta[2]
  b0_1  <- theta[3]
  bb_1  <- theta[4]
  a.tilde0_1 <- theta[5]
  b.a.tilde_1 <- theta[6]
  nu0_1 <- theta[7]
  b.nu_1 <- theta[8]
  b0_2  <- theta[9]
  bb_2  <- theta[10]
  a.tilde0_2 <- theta[11]
  b.a.tilde_2 <- theta[12]
  nu0_2 <- theta[13]
  b.nu_2 <- theta[14]

  LoLi <- sum(
  as.numeric(Status==1)*(
```

```
   ln.Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group) +
        Sat.Ggv.log.dens(b0=b0_1,b1=bb_1,
        a.tilde=a.tilde0_1,b.a.tilde=b.a.tilde_1,
         nu_0=nu0_1,b.nu=b.nu_1,X=Group,ti=Time)) +
  as.numeric(Status==2)*(
   ln.Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group) +
        Sat.Ggv.log.dens(b0=b0_2,b1=bb_2,
         a.tilde=a.tilde0_2,b.a.tilde=b.a.tilde_2,
         nu_0=nu0_2,b.nu=b.nu_2,X=Group,ti=Time)) +
  as.numeric(Status==0)*log(
   Pi_1.x(bpi0=bpi0,bpix=bpix,X=Group) *
        Sat.Ggv.Surv(b0=b0_1,b1=bb_1,a.tilde=a.tilde0_1,
         b.a.tilde=b.a.tilde_1,nu_0=nu0_1,b.nu=b.nu_1,
         X=Group,ti=Time)
        Pi_2.x(bpi0=bpi0,bpix=bpix,X=Group) *
        Sat.Ggv.Surv(b0=b0_2,b1=bb_2,a.tilde=a.tilde0_2,
         b.a.tilde=b.a.tilde_2,nu_0=nu0_2,b.nu=b.nu_2,
         X=Group,ti=Time))
  )
  return(-LoLi)
}

# Log-likelihood for the spline approach
Loglik.Splines <- function(
theta,Time,Status,Group,Basis,D.square,mu)
{
 n.th <- dim(Basis$Basisfkt)[2]
 bpi0 <- theta[1]
 bpi1 <- theta[2]

 eta.pi <- bpi0 + bpi1*Group

 theta.basis.1_A <- theta[3:(2+n.th)]
 theta.basis.1_B <- theta[(3+n.th):(2*n.th+2)]
 theta.basis.2_A <- theta[(3+2*n.th):(3*n.th+2)]
 theta.basis.2_B <- theta[(3+3*n.th):(4*n.th+2)]

 h1.cond_A <- as.vector(
  exp(theta.basis.1_A)%*%t(Basis$Basisfkt))
 h1.cond_B <- as.vector(
  exp(theta.basis.1_A + theta.basis.1_B)%*%t(Basis$Basisfkt))
 h2.cond_A <- as.vector(
  exp(theta.basis.2_A)%*%t(Basis$Basisfkt))
 h2.cond_B <- as.vector(
  exp(theta.basis.2_A + theta.basis.2_B)%*%t(Basis$Basisfkt))

 S1.cond_A <- as.vector(
  exp( - exp(theta.basis.1_A)%*%t(Basis$IFkt-Basis$IB.Min) ) )
 S1.cond_B <- as.vector(
```

```
  exp( - exp(theta.basis.1_A + theta.basis.1_B)%*%
  t(Basis$IFkt-Basis$IB.Min) ) )
 S2.cond_A <- as.vector(
  exp( - exp(theta.basis.2_A)%*%
  t(Basis$IFkt-Basis$IB.Min) ) )
 S2.cond_B <- as.vector(
  exp( - exp(theta.basis.2_A + theta.basis.2_B)%*%
  t(Basis$IFkt-Basis$IB.Min) ) )

 f1.cond_A <- h1.cond_A * S1.cond_A
 f1.cond_B <- h1.cond_B * S1.cond_B
 f2.cond_A <- h2.cond_A * S2.cond_A
 f2.cond_B <- h2.cond_B * S2.cond_B


 loglik <- sum(
     (Status==1)* (eta.pi - log(1+exp(eta.pi)) +
     log(f1.cond_A*(Group==0) + f1.cond_B*(Group==1) ) ) +
        (Status==2)* ( - log(1+exp(eta.pi)) + log(
        f2.cond_A*(Group==0) + f2.cond_B*(Group==1) ) ) +
     (Status==0)*  log( exp(eta.pi)/(1+exp(eta.pi)) *
      (S1.cond_A * (Group==0) + S1.cond_B * (Group==1)) +
        ( 1/(1+exp(eta.pi)) * (S2.cond_A * (Group==0) +
         S2.cond_B * (Group==1) )) ) )

  Pen.Mat <- t(theta) %*% D.square %*% theta
   return( - loglik + 1/2*mu*Pen.Mat)
}
```

Density and survivor functions

Code for density and survivor functions, used for estimation of marginal event type distributions and conditional event time distributions, is presented.

```
# Maringal event type distribution:
Pi_1.x <- function(bpi0,bpix,X)
exp(bpi0 + bpix*X) / (1 + exp(bpi0 + bpix*X))

# Function for P_2(X):
Pi_2.x <- function(bpi0,bpix,X)
1  / (1 + exp(bpi0 + bpix*X))

# Function for ln(P_1(X)):
ln.Pi_1.x <- function(bpi0,bpix,X)
bpi0 + bpix*X - log(1+exp(bpi0 + bpix*X))
```

```
# Function for ln(P_2(X)):
ln.Pi_2.x <- function(bpi0,bpix,X)
- log(1+exp(bpi0 + bpix*X))


################################################
# EXPONENTIAL DISTRIBUTION
################################################

# Density Function
Expo.dens <- function(b0,b1,X,ti)
exp(b0+b1*X) * exp(-exp(b0+b1*X)*ti)

# Logarithm of the Density Function
Expo.log.dens <- function(b0,b1,X,ti)
b0+b1*X - exp(b0+b1*X)*ti

# Survivor Function
Expo.Surv <- function(b0,b1,X,ti)
exp(-exp(b0+b1*X)*ti)


################################################
# WEIBULL DISTRIBUTION
################################################

# Density Function
Weibull.dens <- function(b0,b1,alpha,X,ti)
exp(b0+b1*X) * alpha * (exp(b0+b1*X)*ti)^(alpha-1) *
exp(-(exp(b0+b1*X)*ti)^alpha)

# Logarithm of the Density Function
Weibull.log.dens <- function(b0,b1,alpha,X,ti)
( b0+b1*X) + log(alpha) + (alpha-1)*((b0+b1*X)+log(ti)) -
(exp(b0+b1*X)*ti)^alpha

# Survivor Function
Weibull.Surv <- function(b0,b1,alpha,X,ti)
exp(-(exp(b0+b1*X)*ti)^alpha)


################################################
# GENERALIZED GAMMA DISTRIBUTION
################################################

# Density Function
Ggv.dens <- function(b0,b1,a.tilde,nu,X,ti)
abs(nu) / (a.tilde*ti*gamma(nu^(-2))) * (nu^(-2))*
(exp(b0+b1*X)*ti)^(nu/a.tilde))^(nu^(-2)) *
```

```
    exp(-nu^(-2)*(exp(b0+b1*X)*ti)^(nu/a.tilde))

# Logarithm of the Density Function
Ggv.log.dens <- function(b0,b1,a.tilde,nu,X,ti)
log(abs(nu)) - log(a.tilde) - log(ti) - log(gamma(nu^(-2))) +
    nu^(-2)*(log(nu^(-2)) + nu/a.tilde * (b0+b1*X+log(ti))) -
    nu^(-2)*(exp(b0+b1*X)*ti)^(nu/a.tilde)

# Survivor Function
Ggv.Surv <- function(b0,b1,a.tilde,nu,X,ti)
as.numeric(nu>0) * (1 - pgamma(nu^(-2)*
(exp(b0+b1*X)*ti)^(nu/a.tilde),nu^(-2))) +
   as.numeric(nu<0) * pgamma(nu^(-2)*
   (exp(b0+b1*X)*ti)^(nu/a.tilde),nu^(-2))


################################################
# SATURATED GENERALIZED GAMMA DISTRIBUTION
################################################

# Density Function
Sat.Ggv.dens <- function(
b0,b1,a.tilde_0,b.a.tilde,nu_0,b.nu,X,ti)
abs(nu_0+b.nu*X) / ((a.tilde_0+b.a.tilde*X)*ti*
gamma((nu_0+b.nu*X)^(-2))) * ((nu_0+b.nu*X)^(-2)*
(exp(b0+b1*X)*ti)^((nu_0+b.nu*X)/
(a.tilde_0+b.a.tilde*X)))^((nu_0+b.nu*X)^(-2)) *
    exp(-(nu_0+b.nu*X)^(-2)*
    (exp(b0+b1*X)*ti)^((nu_0+b.nu*X)/(a.tilde_0+b.a.tilde*X)) )

# Logarithm of the Density Function
Sat.Ggv.log.dens <- function(
b0,b1,a.tilde_0,b.a.tilde,nu_0,b.nu,X,ti)
log(abs(nu_0+b.nu*X)) - log(a.tilde_0+b.a.tilde*X) -
log(ti) - log(gamma((nu_0+b.nu*X)^(-2))) +
    (nu_0+b.nu*X)^(-2)*(log((nu_0+b.nu*X)^(-2)) +
    (nu_0+b.nu*X)/(a.tilde_0+b.a.tilde*X) *
    ((b0+b1*X)+log(ti))) - (nu_0+b.nu*X)^(-2)*
    (exp(b0+b1*X)*ti)^((nu_0+b.nu*X)/(a.tilde_0+b.a.tilde*X))

# Survivor Function
Sat.Ggv.Surv <- function(
b0,b1,a.tilde_0,b.a.tilde,nu_0,b.nu,X,ti)
as.numeric((nu_0+b.nu*X)>0) * (1 - pgamma((nu_0+b.nu*X)^(-2)*
(exp(b0+b1*X)*ti)^((nu_0+b.nu*X)/(a.tilde_0+b.a.tilde*X)),
(nu_0+b.nu*X)^(-2))) +  as.numeric((nu_0+b.nu*X)<0) *
pgamma((nu_0+b.nu*X)^(-2)*
(exp(b0+b1*X)*ti)^((nu_0+b.nu*X)/(a.tilde_0+b.a.tilde*X)),
(nu_0+b.nu*X)^(-2))
```

```
################################################
# SPLINES
################################################

# Density Function
haz.Splines <- function(Basis,b.spl)
as.vector(exp(b.spl) %*% t(Basis$Basisfkt))

# Survivor Function
Surv.Splines <- function(Basis,b.spl)
as.vector(exp(-exp(b.spl) %*% t(Basis$IFkt - Basis$IB.Min)))
```

B-spline basis functions and penalty matrix

Definition of basis functions and of the penalty matrix, used for smooth estimation of the conditional hazard rates.

```
################################################
# Splines
################################################

# Basis functions
Bsplines.quant <- function(
TimeEval,TimeKnots,Status,n.Knots,IB=TRUE)
{
# Definition of knots
 Knots <- quantile(
  TimeKnots[which(Status!=0)],seq(0,1,length=n.Knots))
 stime <- sort(TimeEval)
 slack.low <- c(Knots[1]-3*(Knots[2]-Knots[1]),
 Knots[1]-2*(Knots[2]-Knots[1]),Knots[1]-(Knots[2]-Knots[1]))
 slack.upp <- c(rev(Knots)[1]+1*(rev(Knots)[1]-rev(Knots)[2]),
 rev(Knots)[1]+2*(rev(Knots)[1]-rev(Knots)[2]),
  rev(Knots)[1]+3*(rev(Knots)[1]-rev(Knots)[2]))
 All.Knots <- c(slack.low,Knots,slack.upp)


 B_k.t <- matrix(
  nrow=length(TimeEval),ncol=length(All.Knots)-4)
 IB_k.t <- matrix(
  nrow=length(TimeEval),ncol=length(All.Knots)-4)


# Definition of basis functions
 for(k in 1:(length(All.Knots)-4))
```

```
 {
  B_k.t[,k] <- (All.Knots[k+4]-All.Knots[k]) *
    ((All.Knots[k]-stime)^3*as.numeric(
    (All.Knots[k]-stime)>0) / ((All.Knots[k+1]-All.Knots[k])*
    (All.Knots[k+2]-All.Knots[k])*
    (All.Knots[k+3]-All.Knots[k])*
    (All.Knots[k+4]-All.Knots[k]))  +
    (All.Knots[k+1]-stime)^3*as.numeric(
    (All.Knots[k+1]-stime)>0) /
    ((All.Knots[k]-All.Knots[k+1])*
    (All.Knots[k+2]-All.Knots[k+1])*
    (All.Knots[k+3]-All.Knots[k+1])*
    (All.Knots[k+4]-All.Knots[k+1]))  +
    (All.Knots[k+2]-stime)^3*as.numeric(
    (All.Knots[k+2]-stime)>0) /
    ((All.Knots[k]-All.Knots[k+2])*
    (All.Knots[k+1]-All.Knots[k+2])*
    (All.Knots[k+3]-All.Knots[k+2])*
    (All.Knots[k+4]-All.Knots[k+2]))  +
    (All.Knots[k+3]-stime)^3*as.numeric(
    (All.Knots[k+3]-stime)>0) /
    ((All.Knots[k]-All.Knots[k+3])*
    (All.Knots[k+1]-All.Knots[k+3])*
    (All.Knots[k+2]-All.Knots[k+3])*
    (All.Knots[k+4]-All.Knots[k+3]))  +
    (All.Knots[k+4]-stime)^3*as.numeric(
    (All.Knots[k+4]-stime)>0) /
    ((All.Knots[k]-All.Knots[k+4])*
    (All.Knots[k+1]-All.Knots[k+4])*
    (All.Knots[k+2]-All.Knots[k+4])*
    (All.Knots[k+3]-All.Knots[k+4])))
  }


# Integral of basis functions
if(IB==T)
{
  for(k in 1:(length(All.Knots)-4))
  {
   IB_k.t[,k] <- -(All.Knots[k+4]-All.Knots[k])/4 * (
       (All.Knots[k]-stime)^4*as.numeric(
       (All.Knots[k]-stime)>0) /
       ( (All.Knots[k+1]-All.Knots[k])*(All.Knots[k+2]-
       All.Knots[k])*(All.Knots[k+3]-All.Knots[k])*
       (All.Knots[k+4]-All.Knots[k])) +
       (All.Knots[k+1]-stime)^4*as.numeric(
       (All.Knots[k+1]-stime)>0) /
       ((All.Knots[k]-All.Knots[k+1])*
       (All.Knots[k+2]-All.Knots[k+1])*
```

```
      (All.Knots[k+3]-All.Knots[k+1])*
      (All.Knots[k+4]-All.Knots[k+1])) +
      (All.Knots[k+2]-stime)^4*as.numeric(
      (All.Knots[k+2]-stime)>0) /
      ((All.Knots[k]-All.Knots[k+2])*
      (All.Knots[k+1]-All.Knots[k+2])*
      (All.Knots[k+3]-All.Knots[k+2])*
      (All.Knots[k+4]-All.Knots[k+2])) +
      (All.Knots[k+3]-stime)^4*as.numeric(
      (All.Knots[k+3]-stime)>0) /
      ((All.Knots[k]-All.Knots[k+3])*
      (All.Knots[k+1]-All.Knots[k+3])*
      (All.Knots[k+2]-All.Knots[k+3])*
      (All.Knots[k+4]-All.Knots[k+3])) +
      (All.Knots[k+4]-stime)^4*as.numeric(
      (All.Knots[k+4]-stime)>0) /
      ((All.Knots[k]-All.Knots[k+4])*
      (All.Knots[k+1]-All.Knots[k+4])*
      (All.Knots[k+2]-All.Knots[k+4])*
      (All.Knots[k+3]-All.Knots[k+4])))
  }


IB.Min <- c()
for(k in 1:(length(All.Knots)-4))
{
  IB.Min[k] <- -(All.Knots[k+4]-All.Knots[k])/4 * (
      (All.Knots[k]-0)^4*as.numeric((All.Knots[k]-0)>0) /
      ((All.Knots[k+1]-All.Knots[k])*(All.Knots[k+2]-
      All.Knots[k])*(All.Knots[k+3]-All.Knots[k])*
      (All.Knots[k+4]-All.Knots[k])) +
      (All.Knots[k+1]-0)^4*as.numeric((All.Knots[k+1]-0)>0) /
      ((All.Knots[k]-All.Knots[k+1])*(All.Knots[k+2]-
      All.Knots[k+1])*(All.Knots[k+3]-All.Knots[k+1])*
      (All.Knots[k+4]-All.Knots[k+1])) +
      (All.Knots[k+2]-0)^4*as.numeric((All.Knots[k+2]-0)>0) /
      ((All.Knots[k]-All.Knots[k+2])*(All.Knots[k+1]-
      All.Knots[k+2])*(All.Knots[k+3]-All.Knots[k+2])*
      All.Knots[k+4]-All.Knots[k+2])) +
      (All.Knots[k+3]-0)^4*as.numeric((All.Knots[k+3]-0)>0) /
      ((All.Knots[k]-All.Knots[k+3])*(All.Knots[k+1]-
      All.Knots[k+3])*(All.Knots[k+2]-All.Knots[k+3])*
      (All.Knots[k+4]-All.Knots[k+3])) +
      (All.Knots[k+4]-0)^4*as.numeric((All.Knots[k+4]-0)>0) /
      ((All.Knots[k]-All.Knots[k+4])*
      (All.Knots[k+1]-All.Knots[k+4])*
      (All.Knots[k+2]-All.Knots[k+4])*
      (All.Knots[k+3]-All.Knots[k+4])))
  }
```

```
   IB.Min.matrix <- matrix(rep(IB.Min,length(TimeEval)),
       byrow=T,nrow=length(TimeEval),ncol=length(IB.Min))
}

return(list(Time=stime,Basisfkt=B_k.t,IFkt=IB_k.t,
   IB.Min=IB.Min.matrix,Knoten=All.Knots))
}



# Penalty matrix
PenMat <- function(N.InnerKnots)
{
n.th <- N.InnerKnots + 4
DM <- diff(diag(n.th), diff = 2)
Mat.part <- t(DM) %*% DM
M0 <- matrix(0,nrow=2,ncol=4*n.th+2)
M1 <- cbind(matrix(0,ncol=2,nrow=n.th),
Mat.part,matrix(0,ncol=3*n.th,nrow=n.th))
M2 <- cbind(matrix(0,ncol=2,nrow=n.th),
matrix(0,ncol=n.th,nrow=n.th),Mat.part,
matrix(0,ncol=2*n.th,nrow=n.th))
M3 <- cbind(matrix(0,ncol=2,nrow=n.th)
matrix(0,ncol=2*n.th,nrow=n.th),Mat.part,
matrix(0,ncol=n.th,nrow=n.th))
M4 <- cbind(matrix(0,ncol=2,nrow=n.th)
matrix(0,ncol=3*n.th,nrow=n.th),Mat.part)

Dsquare <- rbind(M0,M1,M2,M3,M4)
return(Dsquare)
}
```

Functions for maximum likelihood estimation

Functions used for numerical derivation of maximum likelihood estimates.

```
# Exponential mixture model
EST.EXPO <- function(Data,Start.Vals){
tryCatch(
NLM.EXPO<<-nlm(Loglik.Expo,Start.Vals,Time=Data$Time,
Status=Data$Status,Group=Data$Group,iterlim=100000000),
error=function(e)
NLM.EXPO <<- list(estimate=rep(NA,6),code=99,minimum=NA,
terations=NA,gradient=rep(NA,6)) )
return(NLM.EXPO)
}
```

```
# Weibull mixture model
EST.WEIBULL <- function(Data,Start.Vals){
tryCatch(
NLM.WEIBULL<<-nlm(Loglik.Weibull,Start.Vals,Time=Data$Time,
Status=Data$Status,Group=Data$Group,iterlim=100000000),
error=function(e)
NLM.WEIBULL <<- list(estimate=rep(NA,8),code=99,minimum=NA,
iterations=NA,gradient=rep(NA,8)) )
return(NLM.WEIBULL)
}


# Gen. gamma (lambda) mixture model
EST.GGV <- function(Data,Start.Vals){
tryCatch(
NLM.GGV<<-nlm(Loglik.Ggv,Start.Vals,Time=Data$Time,
tatus=Data$Status,Group=Data$Group,iterlim=100000000),
error=function(e)
NLM.GGV <<- list(estimate=rep(NA,10),code=99,minimum=NA,
iterations=NA,gradient=rep(NA,10)) )
return(NLM.GGV)
}


# Saturated gen. gamma mixture model
EST.SAT.GGV <- function(Data,Start.Vals){
tryCatch(
NLM.SAT.GGV<<-nlm(Loglik.Sat.Ggv,Start.Vals,Time=Data$Time,
Status=Data$Status,Group=Data$Group,iterlim=100000000),
error=function(e)
NLM.SAT.GGV <<- list(estimate=rep(NA,14),code=99,minimum=NA,
iterations=NA,gradient=rep(NA,14)) )
return(NLM.SAT.GGV)
}


# Spline mixture model
EST.SPLINE <- function(Data,mu,N.innerKnots,Start.Vals){
tryCatch(
NLM.SPLINE<<-nlm(Loglik.Splines,Start.Vals,Time=Data$Time,
        Status=Data$Status,Group=Data$Group,
        Basis=Bsplines.quant(TimeEval=Data$Time,
        TimeKnots=Data$Time,
        Status=Data$Status,IB=TRUE,n.Knots=N.innerKnots+2),
        D.square=PenMat(N.innerKnots),mu=mu,
        iterlim=100000000),
error=function(e)
```

```
NLM.SPLINE <<- list(estimate=rep(NA,
        2+4*(N.innerKnots+4)),code=99,
        minimum=NA,iterations=NA,
        gradient=rep(NA,2+4*(N.innerKnots+4))) )
return(NLM.SPLINE)
}
```

Estimation of cause-specific and subdistribution hazards from the mixture models

Functions for derivation of cause-specific and subdistribution hazards and hazard ratios are presented for the saturated generalized gamma and the spline approach. For the exponential and the Weibull mixture model cause-specific and subdistribution hazards were calculated accordingly.

```
################################################
# Functions for Estimation of
# Cause-specific and Subdistribution Hazards
################################################

# Saturated gen. gamma mixture model
Deriv.Sat.Ggv <- function(Timepoints,theta)
{
  # Define parameters
  bpi0  <- theta[1]
  bpix  <- theta[2]
  b0_1  <- theta[3]
  bb_1  <- theta[4]
  a.tilde0_1 <- theta[5]
  b.a.tilde_1 <- theta[6]
  nu0_1 <- theta[7]
  b.nu_1 <- theta[8]
  b0_2  <- theta[9]
  bb_2  <- theta[10]
  a.tilde0_2 <- theta[11]
  b.a.tilde_2 <- theta[12]
  nu0_2 <- theta[13]
  b.nu_2 <- theta[14]

  # Event type probabilities
  Pi1_A <- exp(bpi0) / (1+exp(bpi0))
  Pi1_B <- exp(bpi0+bpix) / (1+exp(bpi0+bpix))
  Pi2_A <- 1-Pi1_A
  Pi2_B <- 1-Pi1_B

  # Density functions of conditional event time distributions
  f_1.A <- Sat.Ggv.dens(b0_1,bb_1,a.tilde0_1,
```

```
 b.a.tilde_1,nu0_1,b.nu_1,X=0,ti=Timepoints)
f_1.B <- Sat.Ggv.dens(b0_1,bb_1,a.tilde0_1,
 b.a.tilde_1,nu0_1,b.nu_1,X=1,ti=Timepoints)
f_2.A <- Sat.Ggv.dens(b0_2,bb_2,a.tilde0_2,
 b.a.tilde_2,nu0_2,b.nu_2,X=0,ti=Timepoints)
f_2.B <- Sat.Ggv.dens(b0_2,bb_2,a.tilde0_2,
 b.a.tilde_2,nu0_2,b.nu_2,X=1,ti=Timepoints)

# Survivor functions of conditional event time distributions
S_1.A <- Sat.Ggv.Surv(b0_1,bb_1,a.tilde0_1,
 b.a.tilde_1,nu0_1,b.nu_1,X=0,ti=Timepoints)
S_1.B <- Sat.Ggv.Surv(b0_1,bb_1,a.tilde0_1,
 b.a.tilde_1,nu0_1,b.nu_1,X=1,ti=Timepoints)
S_2.A <- Sat.Ggv.Surv(b0_2,bb_2,a.tilde0_2,
 b.a.tilde_2,nu0_2,b.nu_2,X=0,ti=Timepoints)
S_2.B <- Sat.Ggv.Surv(b0_2,bb_2,a.tilde0_2,
 b.a.tilde_2,nu0_2,b.nu_2,X=1,ti=Timepoints)

# Cumulative incidence functions
sub.F_1.A <- Pi1_A*(1 - S_1.A)
sub.F_1.B <- Pi1_B*(1 - S_1.B)
sub.F_2.A <- Pi2_A*(1 - S_2.A)
sub.F_2.B <- Pi2_B*(1 - S_2.B)

# Subdensity functions
sub.f_1.A <- Pi1_A*f_1.A
sub.f_1.B <- Pi1_B*f_1.B
sub.f_2.A <- Pi2_A*f_2.A
sub.f_2.B <- Pi2_B*f_2.B

# Overall survivor functions
S.overall_A <- Pi1_A*S_1.A + Pi2_A*S_2.A
S.overall_B <- Pi1_B*S_1.B + Pi2_B*S_2.B

# Estimates for cause-specific hazards (k=1)
csh1_A <- sub.f_1.A / S.overall_A
csh1_B <- sub.f_1.B / S.overall_B
csHR_1 <- csh1_B / csh1_A

# Estimates for subdistribution hazards (k=1)
sdh1_A <- sub.f_1.A / (1 - sub.F_1.A)
sdh1_B <- sub.f_1.B / (1 - sub.F_1.B)
sdHR_1 <- sdh1_B / sdh1_A

# Estimates for cause-specific hazards (k=2)
csh2_A <- sub.f_2.A / S.overall_A
csh2_B <- sub.f_2.B / S.overall_B
csHR_2 <- csh2_B / csh2_A
```

```
  # Estimates for subdistribution hazards (k=2)
  sdh2_A <- sub.f_2.A / (1 - sub.F_2.A)
  sdh2_B <- sub.f_2.B / (1 - sub.F_2.B)
  sdHR_2 <- sdh2_B / sdh2_A

 return(data.frame(Time=Timepoints,
  csh1_A=csh1_A,csh1_B=csh1_B,csHR_1=csHR_1,
        sdh1_A=sdh1_A,sdh1_B=sdh1_B,sdHR_1=sdHR_1,
        csh2_A=csh2_A,csh2_B=csh2_B,csHR_2=csHR_2,
        sdh2_A=sdh2_A,sdh2_B=sdh2_B,sdHR_2=sdHR_2 ))
}


# Spline mixture model
Deriv.Splines <- function(Basis,theta)
{
  # Define parameters
  n.th <- dim(Basis$Basisfkt)[2]
  bpi0  <- theta[1]
  bpix  <- theta[2]

  # Event type probabilities
  Pi1_A <- exp(bpi0) / (1+exp(bpi0))
  Pi1_B <- exp(bpi0+bpix) / (1+exp(bpi0+bpix))
  Pi2_A <- 1-Pi1_A
  Pi2_B <- 1-Pi1_B

  # Hazard functions of conditional event time distributions
  h_1.A <- haz.Splines(Basis=Basis,
      b.spl=theta[3:(n.th+2)])
  h_1.B <- haz.Splines(Basis=Basis,
      b.spl=theta[(3:(n.th+2))]+theta[(n.th+3):(2*n.th+2)])
  h_2.A <- haz.Splines(Basis=Basis,
      b.spl=theta[(2*n.th+3):(3*n.th+2)])
  h_2.B <- haz.Splines(Basis=Basis,
       b.spl=theta[(2*n.th+3):(3*n.th+2)]+
        theta[(3*n.th+3):(4*n.th+2)])

  # Survivor functions of conditional event time distributions
  S_1.A <- Surv.Splines(Basis=Basis,
      b.spl=theta[3:(n.th+2)])
  S_1.B <- Surv.Splines(Basis=Basis,
      b.spl=theta[(3:(n.th+2))]+theta[(n.th+3):(2*n.th+2)])
  S_2.A <- Surv.Splines(Basis=Basis,
      b.spl=theta[(2*n.th+3):(3*n.th+2)])
  S_2.B <- Surv.Splines(Basis=Basis,
      b.spl=theta[(2*n.th+3):(3*n.th+2)]+
       theta[(3*n.th+3):(4*n.th+2)])
```

```
 # Calculation of density functions of conditional event
 # time distributions from conditional hazard rates and
 # conditional survivor functions
 f_1.A <- h_1.A * S_1.A
 f_1.B <- h_1.B * S_1.B
 f_2.A <- h_2.A * S_2.A
 f_2.B <- h_2.B * S_2.B

 # Cumulative incidence functions
 sub.F_1.A <- Pi1_A*(1 - S_1.A)
 sub.F_1.B <- Pi1_B*(1 - S_1.B)
 sub.F_2.A <- Pi2_A*(1 - S_2.A)
 sub.F_2.B <- Pi2_B*(1 - S_2.B)

 # Subdensity functions
 sub.f_1.A <- Pi1_A*f_1.A
 sub.f_1.B <- Pi1_B*f_1.B
 sub.f_2.A <- Pi2_A*f_2.A
 sub.f_2.B <- Pi2_B*f_2.B

 # Overall survivor functions
 S.overall_A <- Pi1_A*S_1.A + Pi2_A*S_2.A
 S.overall_B <- Pi1_B*S_1.B + Pi2_B*S_2.B

 # Estimates for cause-specific hazards (k=1)
 csh1_A <- sub.f_1.A / S.overall_A
 csh1_B <- sub.f_1.B / S.overall_B
 csHR_1 <- csh1_B / csh1_A

  # Estimates for subdistribution hazards (k=1)
 sdh1_A <- sub.f_1.A / (1 - sub.F_1.A)
 sdh1_B <- sub.f_1.B / (1 - sub.F_1.B)
 sdHR_1 <- sdh1_B / sdh1_A

 # Estimates for cause-specific hazards (k=2)
 csh2_A <- sub.f_2.A / S.overall_A
 csh2_B <- sub.f_2.B / S.overall_B
 csHR_2 <- csh2_B / csh2_A

 # Estimates for subdistribution hazards (k=2)
 sdh2_A <- sub.f_2.A / (1 - sub.F_2.A)
 sdh2_B <- sub.f_2.B / (1 - sub.F_2.B)
 sdHR_2 <- sdh2_B / sdh2_A

return(data.frame(Time=Basis$Time,
 csh1_A=csh1_A,csh1_B=csh1_B,csHR_1=csHR_1,
       sdh1_A=sdh1_A,sdh1_B=sdh1_B,sdHR_1=sdHR_1,
       csh2_A=csh2_A,csh2_B=csh2_B,csHR_2=csHR_2,
       sdh2_A=sdh2_A,sdh2_B=sdh2_B,sdHR_2=sdHR_2 ))  }
```

Appendix C

Sketch of R-Code used for data analysis

In this section the R-code used for analysis of the data from the clinical cohort study, which is presented in Section 8, is sketched. Variables considered:

- Time: Event time or censoring time
- Status: Indicating type of event or a censored observation

  1 = cardiac death
  2 = non-cardiac
  0 = censored
- Group: Indicating risk group

  0 = low risk group
  1 = high risk group
- Age: Indicating patient's age

  0 = (age<65 years)
  1 = (age≥65 years)
- Diab: Indicating diabetes

  0 = no diabetes
  1 = diabetes

C.1 Cause-specific hazard regression for the event of interest

```
require(survival)
COXcsh <- coxph(Surv(Time,Status==1) ~ Group + Age + Diab)
```

C.2 Subdistribution hazard regression

```
require(cmprsk)
COXsdh <- crr(Time,Status,cbind(Group,Age,Diab),failcode=1)
```

C.3 Mixture model

ECM algorithm as described by Ng and McLachlan (2003) for two possible types of failure and three covariates. For estimation of the cumulative hazard functions a dataset ordered by observed times is required. Expectation and conditional maximization steps have to be iterated until some predefined convergence criterion is fulfilled.

- Expectation for τ_i denoting the probability for a failure of type 1 for individual i in the $(j+1)^{th}$ iteration given j^{th} estimates for μ, $\boldsymbol{\pi}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and the baseline survival functions $S_{01}(t_i, \mathbf{x}_i, \boldsymbol{\beta}_{\mathbf{1}}^{(j)})$ and $S_{02}(t_i, \mathbf{x}_i, \boldsymbol{\beta}_{\mathbf{2}}^{(j)})$ according to Equations (8) and (10) from Ng and McLachlan (2003):

  ```
  p1 <- exp(mu+Group*pi1+Age*pi2+Diab*pi3) /
    (1+exp(mu+Group*pi1+Age*pi2+Diab*pi3))

  p2 <- 1-p1

  tau <- p1*S01^exp(Group*b1_1+Age*b1_2+Diab*b1_3) /
    (p1*S01^exp(Group*b1_1+Age*b1_2+Diab*b1_3) +
    p2*S02^exp(Group*b2_1+Age*b2_2+Diab*b2_3))
  ```

- Function for $(j+1)^{th}$ estimation of μ and $\boldsymbol{\pi}$. *Q0* denotes the logistic regression component of the expectation of the complete-data log-likelihood given the current parameter estimates.

  ```
  require(rootSolve)
  Q0 <- function(MU) {
  mu.opt <- MU[1]
  pi1.opt <- MU[2]
  pi2.opt <- MU[3]
  pi3.opt <- MU[4]
  P1 <- exp(mu.opt+pi1.opt*Group+pi2.opt*Age+pi3.opt*Diab)/
   (1+exp(mu.opt+pi1.opt*Group+pi2.opt*Age+pi3.opt*Diab))
  fct.mu<-sum((
        as.numeric(Status==1)+(Status==0)*tau-P1))
  fct.p1<-sum((
        as.numeric(Status==1)+(Status==0)*tau-P1)*Group)
  fct.p2<-sum((
        as.numeric(Status==1)+(Status==0)*tau-P1)*Age)
  fct.p3<-sum((
  ```

```
      as.numeric(Status==1)+(Status==0)*tau-P1)*Diab)
return(c(fct.mu,fct.p1,fct.p2,fct.p3)) }

opt <- multiroot(Q0,c(0,0,0,0))

mu.new <- opt$root[1]
pi1.new <- opt$root[2]
pi2.new <- opt$root[3]
pi3.new <- opt$root[4]
```

- Calculation of the cumulative baseline hazard function and the baseline survivor function for event type 1 in the $(j+1)^{th}$ iteration according to Equation (12) from Ng and McLachlan (2003). Measures for event type 2 can be estimated analogously:

```
h01 <- c()

# Estimation of baseline hazard rate
n <- length(Time)
for(i in 1:n)
  h01[i] <- 1 / sum((((Status[i:n]==1) +
    (Status[i:n]==0)*tau[i:n])* exp(Group[i:n]*b1_1+
    Age[i:n]*b1_2+Diab[i:n]*b1_3)))*(Status[i]==1)

# Replace empty components at the end with zeros
h01[which(is.na(h01))] <- 0

# Calculate cumulative baseline hazard function
H01 <- cumsum(h01)

# Calculate baseline survival function
S01 <- exp(-H01)
```

- Conditional maximization step to obtain the $(j+1)^{th}$ estimate for $\boldsymbol{\beta}_1$. According to Equation (9) from Ng and McLachlan (2003) maximization can be conducted separately for all types of event. $\boldsymbol{\beta}_2^{(j+1)}$ can be obtained analogously.

```
# Event type 1:
Q1 <- function(b1.opt) {
 b1.opt_1 <- b1.opt[1]
 b1.opt_2 <- b1.opt[2]
 b1.opt_3 <- b1.opt[3]
 eta1 <- Group*b1.opt_1 + Age*b1.opt_2 + Diab*b1.opt_3

 fct1 <- sum((
  (Status==1)-((Status==1)+(Status==0)*tau) *
  H01*exp(Group*b1.opt_1+
```

```
       Age*b1.opt_2+Diab*b1.opt_3))*Group)
 fct2 <- sum((
       (Status==1)-((Status==1)+(Status==0)*tau) *
  H01*exp(Group*b1.opt_1+
       Age*b1.opt_2+Diab*b1.opt_3))*Age)
 fct3 <- sum((
       (Status==1)-((Status==1)+(Status==0)*tau) *
  H01*exp(Group*b1.opt_1+
       Age*b1.opt_2+Diab*b1.opt_3))*Diab)
 return(c(fct1,fct2,fct3)) }

b1 <- multiroot(Q1,c(0,0,0))$root
b1_1.new <- b1[1]
b1_2.new <- b1[2]
b1_3.new <- b1[3]
```

C.4 Vertical Modelling

- Cox regression for marginal event time distribution considering covariates

```
coxph(Surv(Time,Status>=1) ~ Group + Age + Diab)
```

- Estimation of relative hazards from a logistic regression model including B-splines for flexible influence of *Time* and interaction between *Group* and *Time*. Only individuals with an observed event can be considered. Estimates of relative hazards for cardiac and non-cardiac death for the high risk group are calculated from the regression coefficients.

```
library(splines)
GLM <- glm(Status==1 ~ Group * bs(Time) + Diab + Age,
           family=binomial(link="logit"),subset=Status>0)

# Relative hazard for cardiac death
# in the high risk group
rel.haz.highrisk.cardiac <-
 predict(GLM,type="response",newdata=data.frame(Group=1,
 Time=seq(0,5,length=300),Diab=mean(Diab),Age=mean(Age)))

# Relative hazard for non-cardiac death
# in the high risk group
rel.haz.highrisk.noncardiac <- 1 - rel.haz.highrisk.cardiac
```

C.5 Pseudo observations

R code for generation of pseudo observations and estimation of covariate effects applying a GEE model can be found in Klein et al (2008).

C.6 P-spline mixture model approach

Application of the P-spline mixture model approach was performed using the R functions presented in Section B.2.2 for analysis of the simulated data.

Bibliography

Aalen O (1978a) An empirical transition matrix for non-homogeneous markov chains based on censored observations. Scandinavian Journal of Statistics 5:141–150

Aalen O (1978b) Nonparametric estimation of partial transition probabilities in multiple decrement models. Annals of Statistics 6:534–545, DOI 10.1214/aos/1176344198

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723, DOI 10.1109/TAC.1974.1100705

Allignol A (2011) kmi: Kaplan-Meier multiple imputation for the analysis of cumulative incidence functions in the competing risks setting. URL `http://CRAN.R-project.org/package=kmi`, R package version 0.5

Allignol A, Beyersmann J, Schumacher M (2008) mvna: An R package for the Nelson-Aalen estimator in multistate models. R news 8(2):48–50, URL `http://cran.r-project.org/doc/Rnews/Rnews_2008-2.pdf`

Allignol A, Schumacher M, Wanner C, Drechsler C, Beyersmann J (2011) Understanding competing risks: a simulation point of view. BMC medical research methodology 11:86, DOI 10.1186/1471-2288-11-86

Andersen PK, Keiding N (2012) Interpretability and importance of functionals in competing risks and multistate models. Statistics in Medicine 31:1074–1088, DOI 10.1002/sim.4385

Andersen PK, Perme MP (2010) Pseudo-observations in survival analysis. Statistical Methods in Medical Research 19:71–99, DOI 10.1177/0962280209105020

Andersen PK, Klein JP, Rosthøj S (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 90:15–27, DOI 10.1093/biomet/90.1.15

Bakoyannis G, Touloumi G (2011) Practical methods for competing risks data: A review. Statistical Methods in Medical Research 21:257–272, DOI 10.1177/0962280210394479

Barthel P, Schneider R, Bauer A, Ulm K, Schmitt C, Schömig A, Schmidt G (2003) Risk stratification after acute myocardial infarction by heart rate turbulence. Circulation 108:1221–1226, DOI 10.1161/01.CIR.0000088783.34082.89

Bauer A, Kantelhardt JW, Barthel P, Schneider R, Makikallio T, Ulm K, Hnatkova K, Schömig A, Huikuri H, Bunde A, Malik M, Schmidt G (2006) Declaration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. Lancet 367:1674–1681, DOI 10.1016/S0140-6736(06)68735-7

Bauer A, Barthel P, Schneider R, Ulm K, Müller A, Joeining A, Stich R, Kiviniemi A, Hnatkova K, Huikuri H, Schömig A, Malik M, Schmidt G (2009) Improved stratification of autonomic regulation for risk prediction in post-infarction patients with preserved left ventricular function (ISAR-risk). European Heart Journal 30:576–583, DOI 10.1093/eurheartj/ehn540

Belot A, Abrahamowicz M, Remontet L, Giorgi R (2010) Flexible modeling of competing risks in survival analysis. Statistics in medicine 29(23):2453–2468, DOI 10.1002/sim.4005

Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine 24(11):1713–1723, DOI 10.1002/sim.2059

Bernoulli D (1766) Essai d'une nouvelle analyse de la mortalité causée par la petite vérole. Mém. Math. Phys. Acad. Roy. Sci., Paris, (Reprinted in: L.P. Bouckaert, B.L. van der Waerden (Eds.), Die Werke von Daniel Bernoulli, Bd. 2 Analysis und Wahrscheinlichkeitsrechnung, Birkhäuser, Basel, 1982, p. 235. English translation entitled 'An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it' in: L. Bradley, Smallpox Inoculation: An Eighteenth Century Mathematical Controversy, Adult Education Department, Nottingham, 1971, p. 21. Reprinted in: S. Haberman, T.A. Sibbett (Eds.) History of Actuarial Science, vol. VIII, Multiple Decrement and Multiple State Models, William Pickering, London, 1995, p. 1.)

Berry SD, Ngo L, Samelson EJ, Kiel DP (2010) Competing risk of death: An important consideration in studies of older adults. Journal of the American Geriatrics Society 58(4):783–787, DOI 10.1111/j.1532-5415.2010.02767.x

Betensky RA, Schoenfeld DA (2001) Nonparametric estimation in a cure model model with random cure times. Biometrics 57:282–286, DOI 10.1111/j.0006-341X.2001.00282.x

Beyersmann J, Schumacher M (2008) Time-dependent covariates in the proportional subdistribution hazards model for competing risks. Biostatistics 9:765–776, DOI 10.1093/biostatistics/kxn009

Beyersmann J, Schumacher M (2007) Letter to the editor: Misspecified regression model for the subdistribution hazard of a competing risk. Statistics in Medicine 26:1649–1652, DOI 10.1002/sim.2727

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 26:5360–5369, DOI 10.1002/sim.3006

Beyersmann J, Latouche A, Buchholz A, Schumacher M (2009) Simulating competing risks data in survival analysis. Statistics in medicine 28(6):956–971, DOI 10.1002/sim.3516

Beyersmann J, Schumacher M, Allignol A (2012) Competing Risks and Multistate Models with R. Springer, New York

Bodnar E, Blackstone EH (2005) Editorial: An 'actual' problem: Another issue of apples and oranges. The Journal of Heart Valve Disease 14(6)

Bodnar E, Blackstone EH (2006) *In Reponse to:* Editorial: An aActual' problem: Another issue of apples and oranges. The Journal of Heart Valve Disease 15(2)

Bradley L (1971) Smallpox Inoculation: An Eigteenth Century Mathematical Controversy. Adult Education Department of the University of Nottingham

Braun TM, Yuan Z (2007) Comparing the small sample performance of several variance estimators under competing risks. Statistics in Medicine 26:1170–1180, DOI 10.1002/sim.2661

Breslow N (1972) Discussion of the paper by D. R. Cox. Journal of the Royal Statistical Society, Series B 34:216–217

Chappell R (2012) Competing risk analyses: How are they different and why should you care? Clinical Cancer Research 18:2127–2129, DOI 10.1158/1078-0432.CCR-12-0455

Chen YH (2010) Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(2):235–251, DOI 10.1111/j.1467-9868.2009.00734.x

Cox C, Chu H, Schneider MF, Muñoz A (2007) Tutorial in biostatistics: Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine 26:4352–4374, DOI 10.1002/sim.2836

Cox DR (1959) The analysis of exponentially distributed life-times with two types of failure. Journal of the Royal Statistical Society, Series B 21(2):411–421

Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society, Series B 34:187–220, DOI 10.2307/2985181

Crowder MJ (2001) Classical Competing Risks. Chapman & Hall/CRC, Boca Raton

Deslandes E, Chevret S (2010) Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: Application to ICU data. BMC medical research methodology 10(1):69, DOI 10.1186/1471-2288-10-69

Dietz K, Heesterbeek AP (2002) Daniel Bernoulli's epidemiological model revisited. Mathematical Biosciences 180(1-2):1–21, DOI 10.1016/s0025-5564(02)00122-0

Dignam JJ, Kocherginsky MN (2008) Choice and interpretation of statistical tests used when competing risks are present. Journal of Clinical Oncology 26(24):4027–4034, DOI 10.1200/jco.2007.12.9866

Dignam JJ, Zhang Q, Kocherginsky M (2012) The use and interpretation of competing risks regression models. Clinical Cancer Research 18(8):2301–2308, DOI 10.1158/1078-0432.ccr-11-2097

Dixon SN, Darlington GA, Desmond AF (2011) A competing risks model for correlated data based on the subdistribution hazard. Lifetime Data Analysis 17:473–495, DOI 10.1007/s10985-011-9198-9

Eilers PH, Marx BD (1996) Flexible smoothing with b-splines and penalties. Statistical science 11(2):89–102, DOI 10.1214/ss/1038425655

Escarela G, Bowater RJ (2008) Fitting a semi-parametric mixture model for competing risks in survival data. Communication in Statistics - Theory and Methods 37:277–293, DOI 10.1080/03610920701649134

Essler M, Wantke J, Mayer B, Scheidhauer K, Bundschuh R, Haller B, Astner ST, Molls M, Andratschke N (2013) Positron-emission tomography CT to identify local recurrence in stage I lung cancer patients 1 year after stereotactic body radiation therapy. Strahlentherapie und Onkologie 189:495–501, DOI 10.1007/s00066-013-0310-9

Fahrmeir L, Tutz G (2001) Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York

Fahrmeir L, Kneib T, Lang S (2007) Regression: Modelle, Methoden und Anwendungen (Statistik und ihre Anwendungen). Springer, Berlin

Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association 94:496–509, DOI 10.2307/2670170

Friedman M (1982) Piecewise exponential models for survival data with covariates. The Annals of Statistics 10:101–113, DOI 10.1214/aos/1176345693

Gaynor JJ, Feuer EJ, Tan C, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF (1993) On the use of cause-specific failure and conditional failure probabilities: Examples from clinical oncology data. Journal of the American Statistical Association 88:400–409, DOI 10.1080/01621459.1993.10476289

Gentle JE (2003) Random Number Generation and Monte Carlo Methods (Statistics and Computing), 2nd edn. Springer, New York

Geskus RB (2011) Cause-specific cumulativei Incidence estimation and the Fine and Gray model under both left truncation and right censoring. Biometrics 67(4):39–49, DOI 10.1111/j.1541-0420.2010.01420.x

Gomes O, Combes C, Dussauchoy A (2008) Parameter estimation of the generalized gamma distribution. Mathematics and Computers in Simulation 79(4):955–963, DOI 10.1016/j.matcom.2008.02.006

Grambauer N, Schumacher M, Beyersmann J (2010) Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. Statistics in Medicine 29:875–884, DOI 10.1002/sim.3786

Gray B (2010) cmprsk: Subdistribution analysis of competing risks. URL `http://CRAN.R-project.org/package=cmprsk`, R package version 2.2-1

Gray R (1988) A class of k-sample tests for comparing the cumulative incidence function in the presence of a competing risk. The Annals of Statistics 16:1141–1154, DOI 10.2307/2241622

Grunkemeier GL, Takkenberg JM, Jamieson WR, Miller DC (2006) *Letter to the Editor - in Response to:* Editorial: An 'actual' problem another issue of apples and oranges. The Journal of Heart Valve Disease 15(2)

Grunkemeier GL, Jin R, Eijkemans MJ, Takkenberg JM (2007) Actual and actuarial probabilities of competing risks: Apples and lemons. The Annals of Thoracic Surgery 83:1586–1592, DOI 0.1016/j.athoracsur.2006.11.044

Haller B, Ulm K (2013) Flexible simulation of competing risks data following prespecified subdistribution hazards. Journal of Statistical Computation and Simulation (ahead-of-print), DOI 10.1080/00949655.2013.793345

Haller B, Schmidt G, Ulm K (2013) Applying competing risks regression models: An overview. Lifetime Data Analysis 19:33–58, DOI 10.1007/s10985-012-9230-8

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer, New York

Hernandez-Quintero A, Dupuy JF, Escarela G (2011) Analysis of a semiparametric mixture model for competing risks. Annals of the Institute of Statistical Mathematics 63(2):305–329, DOI 10.1007/s10463-009-0229-1

Hjort NL (1992) On inference in parametric survival data models. International Statistical Review 60:355–387

Højsgaard S, Halekoh U, J Y (2005) The R package geepack for generalized estimating equations 15:1–11, URL `http://CRAN.R-project.org/package=survival`, R package version 2.36-5

Kaishev VK, Dimitrova DS, Haberman S (2007) Modelling the joint distribution of competing risks survival times using copula functions. Insurance: Mathematics and Economics 41:339–361, DOI 10.1016/j.insmatheco.2006.11.006

Kalbfleisch JD, Prentice RL (2002) The Statistical Analysis of Failure Time Data. John Wiley & Sons, Hoboken, NJ

Kaplan EL, Meier P (1958) Non-parametric estimation from incomplete observations. Journal of the American Statistical Association 53:457–481, DOI 10.1080/01621459.1958.10501452

Katsahian S, Resche-Rigon M, Chevret S, Porcher R (2006) Anaylsing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. Statistics in Medicine (25):4267–4278, DOI 10.1002/sim.2684

Kim HT (2006) Cumulative incidence in competing risks data and competing risks regression analysis. Clinical Cancer Research (13):559–565, DOI 10.1158/1078-0432.CCR-06-1210

Klein J, Gerster M, Andersen P, Tarima S, Perme M (2008) SAS and R functions to compute pseudo-values for censored data regression. Computer Methods and Programs in Biomedicine 89(3):289–300, DOI 10.1016/j.cmpb.2007.11.017

Klein JP (2010) Competing risks. WIREs Computational Statistics 2:333–339, DOI 10.1002/wics.83

Klein JP, Andersen PK (2005) Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 61:223–229, DOI 10.1111/j.0006-341X.2005.031209.x

Klein JP, Moeschberger ML (1988) Bounds on net survival probabilities for dependent competing risks. Biometrics 44:529–539

Klein JP, Moeschberger ML (2003) Survival Analysis - Techniques for Censored and Truncated Data. Springer, New York

Koller MT, Raatz H, Steyerberg EW, Wolbers M (2012) Competing risks and the clinical community: irrelevance or ignorance. Statistics in Medicine 31:1089–1097, DOI 10.1002/sim.4348

Kooperberg CP, Stone CJ, Truong YK (1995) Hazard regression. Journal of the American Statistical Association 90:78–94, DOI 10.1080/01621459.1995.10476491

Korn EL, Dorey FJ (1992) Application of crude incidence curves. Statsitics in Medicine 11:813–829, DOI 10.1002/sim.4780110611

Kuk AY, Chen CH (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79(3):531–541, DOI 10.1093/biomet/79.3.531

Lambert PC, Dickman PW, Nelson CP, Royston P (2010) Estimating the crude probability of death due to cancer and other causes using relative survival models. Statistics in Medicine 29:885–895, DOI 10.1002/sim.3762

Larson MG, Dinse GE (1985) A mixture model for the regression analysis of competing risks data. Journal of the Royal Statistical Society Series C (Applied Statistics) 34:201–211

Latouche A, Porcher R, Chevret S (2004) Sample size formula for proportional hazards modelling of competing risks. Statistics in medicine 23(21):3263–3274, DOI 10.1002/sim.1915

Latouche A, Porcher R, Chevret S (2005) A note on including time-dependent covariate in regression model for competing risks data. Biometrical Journal 47(6):807–814, DOI 10.1002/bimj.200410152

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):965–974, DOI 10.1002/sim.2600

Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP (2013) A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. Journal of Clinical Epidemiology 66(6):648–653, DOI 10.1016/j.jclinepi.2012.09.017

Lau B, Cole SR, Moore SR, Gange SJ (2008) Evaluating competing adverse and beneficial outcomes using a mixture model. Statistics in Medicine 27:4313–4327, DOI 10.1002/sim.3293

Lau B, Cole SR, J GS (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 30:654–665, DOI 10.1002/sim.4123

Leemis LM (1987) Variate generation for accelerated life and proportional hazards models. Operations research 35(6):892–894

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22, DOI 10.1093/biomet/73.1.13

Lin D (1997) Non-parametric inference for cumulative incidence functions in competing risks studies. Statistics in Medicine 16:901–910

Lo SM, Wilke RA (2010) A copula model for dependent competing risks. Journal of the Royal Stistical Society: Series C (Applied Statistics) 59:359–376, DOI 10.1111/j.1467-9876.2009.00695.x

Lunn M, McNeil D (1995) Applying cox regression to competing risks. Biometrics 51:524–532, DOI 10.2307/2532940

Marubini E, Valsecchi MG (1995) Analysing Survival Data from Clinical Trials and Observational Studies. John Wiley & Sons, Chichester

Miller RG (1974) The jackknife - a review. Biometrika 61(1):1–15, DOI 10.1093/biomet/61.1.1

Moeschberger ML, Klein JP (1995) Statistical methods for dependent competing risks. Lifetaime Data Analysis 1:195–204, DOI 10.1007/BF00985770

Nelson W (1969) Hazard plotting for incomplete failure data. Journal of Quality Technology 1:27–52

Ng GK, McLachlan GJ (2003) An EM-based semi-parametric mixture model approach to the regression analysis of competing risks data. Statistics in Medicine 22:1097–1111, DOI 10.1002/sim.1371

Nicolaie MA, van Houwelingen HC, Putter H (2010) Vertical modeling: A pattern mixture approach for competing risks modeling. Statistics in Medicine 29:1190,1205, DOI 10.1002/sim.3844

Noufaily A, Jones MC (2013) On maximization of the likelihood for the generalized gamma distribution. Computational Statistics 28(2):505–517, DOI 10.1007/s00180-012-0314-4

Pepe MS (1991) Inference for events with dependent risks in multiple endpoint studies. Journal of the American Statistical Association 86:770–778, DOI 10.1080/01621459.1991.10475108

Perme MP, Andersen PK (2008) Checking hazard regression models using pseudo-observations. Statistics in Medicine 27:5309–5328, DOI 10.1002/sim.3401

Peterson AV (1976) Bounds for a joint distribution function with fixed sub-distribution fnuctions: Applications to competing risks. Proceedings of the National Academy of Sciences 73:11–13

Pintilie M (2006) Competing risks: A practical perspective. John Wiley & Sons, Chichester, West Sussex

Pintilie M (2007) Analysing and interpreting competing risk data. Statistics in Medicine 26:1360–1367, DOI 10.1002/sim.2655

Prentice R, Kalbfleisch J, Peterson A, Flournoy N, Farewell V, Breslow N (1978) The analysis of failure times in the presence of competing risks. Biometrics 34:541–554, DOI 10.2307/2530374

Putter H, Fiocco M, Geskus RB (2007) Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine 26(11):2389–2430, DOI 10.1002/sim.2712

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0

Resche-Rigon M, Chevret S (2006) Local influence for the subdistribution of a competing risk. Statistics in Medicine 25:1937–1947, DOI 10.1002/sim.2354

Robins J, Rotnitzky A (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K, Farewell V (eds) AIDS Epidemiology - Methodological Issues, Birkhäuser, Boston, pp 24–33

Robins JM, Finkelstein DM (2000) Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 56(3):779–788, DOI 10.1111/j.0006-341X.2000.00779.x

Roobol M, Heinsdijk EA (2011) Propensity score matching, competing risks analysis, and a competing risk nomogram: Some guidance for urologists may be in place. European Urology 60(5):931–933, DOI 10.1016/j.eururo.2011.07.039

Rosenberg PS (1995) Hazard function estimation using B-Splines. Biometrics 51:874–887

Royston P, Parmar MK (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 21:2175–2197, DOI 10.1002/sim.1203

Ruan PK, Gray RJ (2008) Analyses of cumulative incidence functions via non-parametric multiple imuputation. Statistics in Medicine (27):5709–5724, DOI 10.1002/sim.3402

Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD (2004) A note on competing risks in survival analysis. British Journal of Cancer 91:1229–1235, DOI 10.1038/sj.bjc.6602102

Scheike TH, Martinussen T (2006) Dynamic Regression Models for Survival Data. Springer, New York

Scheike TH, Zhang MJJ (2011) Analyzing competing risk data using the R timereg package. Journal of Statistical Software 38:1–15

Schemper M, Smith T (1996) A note on quantifying follow-up in studies of failure time. Lancet 17:343–346, DOI 0.1016/0197-2456(96)00075-x

Schemper M, Wakounig S, Heinze G (2009) The estimation of average hazard ratios by weighted cox regression. Statistics in Medicine 28:2473–2489, DOI 10.1002/sim.3623

Schoenfeld D (1982) Partial residuals for the proportional hazards regression model. Biometrika 69(1):239–241, DOI 10.1093/biomet/69.1.239

Scrucca L, Santucci A, Aversa F (2007) Competing risk analysis using R: An easy guide for clinicians. Bone Marrow Transplantation 40:381–387, DOI 10.1038/sj.bmt.1705727

Slud EV, Rubinstein LV (1983) Dependent competing risks and summary survival curves. Biometrika 70:643–649, DOI 10.1093/biomet/70.3.643

Sun Y, Hyun S, Gilbert P (2008) Testing and estimation of time-varying cause-specific hazard ratios with covariate adjustment. Biometrics 64:1070–1079, DOI 10.1111/j.1541-0420.2008.01012.x

Sylvestre MP, Abrahamowicz M (2008) Comparison of algorithms to generate event times conditional on time-dependent covariates. Statistics in Medicine 27(14):2618–2634, DOI 10.1002/sim.3092

Therneau T (2011) survival: Survival analysis, including penalised likelihood. URL `http://CRAN.R-project.org/package=survival`, R package version 2.36-5

Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model (Statistics for Biology and Health). Springer, New York

Tsiatis A (1975) A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences 72(1):20–22

Tsiatis AA (2005) Competing risks. In: Armitage P, Colton T (eds) Encyclopedia of Biostatistics (second edition), John Wiley & Sons, New York, pp 824–835

Wegman EJ, Wright IW (1983) Splines in statistics. Journal of the American Statistical Association 78(382):351–365, DOI 10.1080/01621459.1983.10477977

Zhang X, Zhang MJJ, Fine J (2011) A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. Statistics in Medicine 30(16):1933–1951, DOI 10.1002/sim.4264

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Haller, Bernhard

Name, Vorname

München, 04.02.2014

Ort, Datum

Unterschrift Doktorand/in