Flexible estimation of complex effects in the context of competing risks survival analysis: Exposure-lag-response association of artificial nutrition and patients' length of stay in intensive care units.

Philipp Kopper

Supervised by: Fabian Scheipl
Submitted in accordance with the requirements for the degree of Master of Science
Institute of Statistics Ludwig Maximilian University Munich

28.04.2020

Contents

List of Tables

List of Figures

Abstract

Often, critically-ill patients in the intensive care unit (ICU) require artificial nutrition. To study the relationship between nutrition and survival, data of ca. 16000 patients from hundreds of ICUs worldwide was collected. For each patient, received calories and proteins (vs. prescribed) were recorded for maximally twelve days.

In an earlier study, piece-wise exponential additive mixed models (PAMMs) are presented to estimate the complex, lagged, cumulative (functional) effects of time-dependent covariates and applied to the nutrition in the ICU (Bender et al. (2018b)). However, Bender et al. (2018b) treat the potential competing risk "hospital discharge" as an administrative censoring event.

In this study, we extend the methodological foundations of PAMMs to the competing risks setting via the

(a) cause-specific hazards (see e.g. Kalbfleisch and Prentice (2011)) and

(b) subdistribution hazards (Fine and Gray (1999)).

Both methods are applied to reevaluate the association between nutrition and length of hospital stay. As Bender et al. (2018b) only model subdistribution hazards, the focus of this thesis will lie on the modeling of cause-specific hazards. Bender et al. (2018b) focus on the effect of caloric intake. This data analysis additionally assesses the effect of protein intake. The outlined methods are implemented in the R package pammtools.

This thesis presents newly developed methods for the modeling of competing risks. Furthermore, it applies these methods to the relevant medical research question described above. Next to the new modeling approach it also uses new data for the same research question. Last, it evaluates in a more general fashion which of the presented models may be favourable.

Chapter 1

Introduction

Bender et al. (2018a) propose the use of piece-wise exponential models (PEM) to flexibly estimate a large number of complex effects in time-to-event analyses. In particular, they focus on the estimation of cumulative effects or exposure-lagged-response associations (ELRA) (Bender et al. (2018b)). In Bender et al. (2018b) and Hartl et al. (2019) this framework is used to assess the impact of nutritional intake (artificial nutrition) on patient survival in intensive care units. Bender et al. (2018b) use data which features the survival time of approx. 10000 patients, information on their diet, and additional health record related information. The main association of interest is the lagged cumulative effect of caloric intake within a twelve-days window on 30-days survival. In sum, they find that a diet that results in a caloric intake between 30 to 70 percent of prescribed calories is associated with significantly increased survival times of critically-ill patients.

Bender et al. (2018a) further generalise PEMs (Cox (1972), Friedman et al. (1982)) to piece-wise additive mixed models (PAMMS). PAMMs are essentially GAMMs with a special induced data structure. PAMMs can virtually model any association that can be modeled by GAMMs. As typically complex effects are easier to express in a GAMM framework than in classical survival models (e.g. Cox proportional hazards model) and there is vast recent research on complex effect estimation in GAMMs, PAMMs can be attractive tools for survival analysis.

PAMMs being rather new, they still lack some generalisation. For example, there has not been any attempt yet to apply PAMMs to competing risks. Equivalently, Bender et al. (2018b) ignore potential competing risks to *death* within their analysis. They treat the (potentially) competing risk of hospital discharge as an administrative censoring event.

The objective of this thesis is to further generalise PAMMs to competing risks PAMMs. The new methodology and its implementation in the Rpackage `pammtools` (Bender and Scheipl (2018)) are presented in this thesis. For both dominant models in competing risk analysis, the cause-specific hazards model, and the subdistribution hazards model (Fine and Gray (1999)), we present PAMM analoga. In our data analysis, the new methodology is used to generalise the approach of Bender et al. (2018b) to competing risks (death vs. discharge) analysis. Our analysis highlights two further significant aspects:

1) Next to the analysis of the caloric intake, we also investigate the effect of protein intake.

2) We use an updated version of the data set used in Bender et al. (2018b) with additional data. Precisely, data from the years 2013 and 2014 are also available now.

Our thesis generates two novel insights. First of all, it presents new techniques for the modeling of competing risks. We develop a complete framework for the analysis of competing risks and embed this into the `pammtools` package. Using the newly presented method one can flexibly model cause-specific and subdistribution hazards.

Second, our data analysis generates clinical insights. In sum, we agree with earlier studies of the data that higher levels of caloric intake are preferable compared to lower ones. Nevertheless, at some point, the effect of additional caloric intake diminishes. our data suggest a diet that supplies between 30 and 70 percent of prescribed calories to be recommendable. Furthermore, we outline that also higher amounts of protein tend to be associated with better outcomes in patients. We find that a diet which administers 0.6 g to 1.2 g protein per Kg body weight in the first days after admission is associated with desired outcomes in patients. A lifting of the protein intake to levels higher than 1.2 g for later days after admission increases these effects. The diet just described is associated with a decreased hospital stay and increased survival time of patients.

The thesis is structured in chapters using the `bookdown` package (Xie (2016)) in `R`. In cases considered beneficial, `R` code is provided. The style of the thesis varies across the chapters. This is due to the different focuses of the chapters. We present theory, software, and application in this thesis. The change of style aims to accommodate this variety of topics. The thesis is structured as follows. Chapter 2 introduces piece-wise additive models using Bender et al. (2018a), Bender (2018) and Bender and Scheipl (2018). Further, chapter 3 reviews the study of Bender et al. (2018b) as baseline literature. A special emphasis will be placed on lagged cumulative effect estimation. Chapter 4 introduces different concepts for competing risks survival analysis. Especially the cause-specific hazards model and the subdistribution hazards model are the central topic of this chapter. Moreover, chapter 5 combines the ideas from the previous chapters (PAMMs and competing risks) to generate a set of new models. At some points, the chapter has the character of a software tutorial. In chapter 6 the new methodology is tested using simulated competing risks data. Subsequently, chapter 7 features our data analysis building on Bender et al. (2018b). Lastly, chapter 8 concludes by reviewing the methodological advances proposed by this thesis and the statistical and clinical findings of the data analysis. The main objective of this thesis is to develop methods for the application in chapter 7. Hence, this analysis is the focus of this thesis.

Chapter 2

A generalised additive model approach to time-to-event analysis

This chapter is mostly based on Bender and Scheipl (2018), Bender et al. (2018a), and their supporting material (such as vignettes). On the one hand, PAMMs are introduced formally by deriving them as a generalisation of an equivalent representation of the Cox PH model. On the other hand, we explicitly provide the implementation in `pammtools` in this chapter.

2.1 Definitions

Within this section, various models will be discussed. We will discuss:

a)

- generalised linear models (Nelder and Wedderburn (1972)) (GLM)
- generalised linear mixed models (McCulloch and Neuhaus (2005)) (GLMM)
- generalised additive models (Hastie (2017)) (GAM)
- generalised additive mixed models (Wood (2017)) (GAMM)

b)

- piece-wise exponential models (Friedman et al. (1982)) (PEM)
- piece-wise exponential mixed models (PEMM)
- piece-wise additive models (PAM)
- piece-wise additive mixed models (Bender et al. (2018a)) (PAMM)

GAMMs are the most general case of the category a) and hence GAMMs will further represent all less general cases (GLM, GAM, GLMM) as well. Equivalently, PAMMs represent – as the most general case – all less general cases in category b) (PEM, PEMM, PAM). PAMMs – as the most general case – will for convenience represent all less general model classes.

2.2 Piece-wise additive mixed models

Piece-wise additive mixed models (PAMMs) (Bender et al. (2018a)) are very useful vehicles for survival analysis. PAMMs are the generalisation of piece-wise exponential models (PEMs) (Cox (1972)). PEMs result from a different representation of the Cox model. The likelihood of a PEM is (under certain assumptions) proportional to the partial likelihood of the corresponding Cox model. Hence, maximum likelihood estimators are equivalent.

However – due to the alternative representation – PEMs work differently from the Cox model. When modeling PEMs (or in the more general case PAMMs) one partitions the follow-up time into intervals. Within these, the hazard rates are assumed to be piece-wise constant. The baseline hazard is not modeled directly but rather the number of events in each interval. This means that PEMs or PAMMs can be modeled as an ordinary generalised linear model (GLM) or generalised additive (mixed) model (GA(M)M) with Poisson likelihood function (Bender et al. (2018a)).

While PEMs are attractive as they can be equivalent to the Cox PH model, we will be more interested in their generalisation, PAMMs. This is because PAMMs – being the most flexible model set up – allow the modeling of a wide range of effects and even a smooth baseline hazard function.

To be able to deal with PAMMs as ordinary GAMMs, data preprocessing is necessary. The preprocessing induces an artificial long format. The follow-up is divided into K mutually exclusive and exhaustive intervals concerning the survival time. This requires $K + 1$ cutting points. For equivalence between the Cox model and PAMMs, these cut points need to equal to the distinct observed event times (Bender (2018)). For each individual, each covariate is *evaluated* at each of the pre-defined cut points. Also, the status variable is always reported at its current state in each interval. In `R` this preprocessing framework is supplied by the `pammtools` package (Bender and Scheipl (2018)). Within this chapter, we explain the preprocessing with the assistance of this package in greater detail. Furthermore, we discuss the properties of this PEM data structure in this section.

2.2.1 Equivalent Poisson model

The Cox proportional hazards (Cox PH) model (Cox (1972)) is the standard model for survival analysis. The following section aims to motivate PEMs to be equivalent to the Cox PH model.

The classical definition of the Cox PH model for a single hazard looks like the following:

$$\lambda_i(t|x_i) = \lambda_0(t) \exp(x_i^T \beta) \tag{2.1}$$

where x_i is a single observational unit. The Cox model estimates the regression coefficients (β) using the partial likelihood. The baseline hazard $\lambda_0(t)$ is estimated non-parametrically for the Cox model using the Nelson-Aalen estimator. If we assume piece-wise constant hazards instead of a fixed baseline hazard the previous definition of the Cox model can be generalised to:

$$\lambda_i(t|x_i) = \lambda_j(t) \exp(x_i^T \beta) \tag{2.2}$$

where j indicates some – so far arbitrary – intervals.

Friedman et al. (1982) show that this model has proportional (partial) likelihood to the following poisson regression model when using a **special** data structure:

$$\lambda_i(t|x_i) = \exp(\log(\lambda_j) + x_i^T\beta + o_{ij}) \tag{2.3}$$

By special, we mean in detail the PEM data structure where the interval cut points equal the (distinct) observed event times.

This alternative model refers to a Poisson regression model with suitable offsets o_{ij}. These offsets reflect additional information on the observed survival time which is not featured by a model without an offset.

Both likelihoods being proportional means that their ML-estimators are equivalent.

2.2.2 A smooth baseline

While this equivalence is the formal motivation of PEMs, the need for a parametric estimation of the baseline is a major shortcoming of the model. Computationally, it is much more expensive than the non-parametric Cox model approach and typically the selection of cut points is somewhat arbitrary (if one deviates from the equivalent representation) (Bender (2018)). Bender et al. (2018a) suggest the use of a smooth baseline. This is facilitated by the semi-parametric smooth estimation of a GAM. A smooth baseline is more agnostic about the cut point selection and (if penalised) and discards this problem greatly. Furthermore, a smooth baseline is a nice model property. The non-parametric estimation of the baseline in the Cox model operates in discrete time – which is typically not how one thinks about risk processes.

2.2.3 The benefits of modeling PAMMs

While PAMMs seem attractive for the straight-forward implementation of complex effects, the Cox model has been the dominating model in survival analysis. This is because only the partial likelihood needs to be optimised in a Cox model making it computationally more efficient. Additionally, the proper specification of a PEM is not trivial since cut point selection may be very arbitrary. A semi-parametric estimation makes cut point selection not a problem anymore. The computational aspects need to be put into context. The PEM representation "blows" up a data set. For static covariate effects, this means that covariate values are often replicated for the estimation. Thus, effectively a much bigger data set is used for estimation. However, if time-dependent covariates enter the study, the differences in the data set complexity assimilate. Still, typically Cox models have an edge on speed (Bender (2018)). At the same time GAMMs – being very general – find application in numerous settings next to survival analysis. Thus, there is more research on effect estimation in the GAMM context. Especially complex effects and the use of mixed model effects are well studied for these models. For example, `mgcv` (Wood (2001)) is **the** standard resource for complex effect modeling in `R`. Using `mgcv` one can easily model smooth effects, time-varying effects, or cumulative effects (e.g. Bender et al. (2018b)).

By modeling PAMMs, one utilises the advances of GAMM research for survival analysis.

2.2.4 Data preprocessing

As outlined in the previous section, the PEM data structure deals with the survival analysis problem as a count data problem. The main feature of this representation is the partition of the follow-up time into a finite number of intervals. Within these intervals hazard rates are assumed to be piece-wise constant.

2.2.4.1 Survival times & events

Below, we show how to transform regular survival data into PEM format where the intervals are equidistant and of size 0.4. Still, it should be chosen with great care. The data set contains a case `id`, a variable indicating the observed event time (`obs_times`) and the observed event (`status`) As usual in survival analysis, the `status` 1 refers to the event while 0 to a censoring. A priori, the selected interval cut points defining the baseline hazard in each interval are arbitrary.

```r
library(pammtools)
library(tidyr)
library(dplyr)
data_1 <- data.frame(id = 1:3, obs_times = c(1, 0.5, 2), status = c(0, 1, 1))
ped_1 <- as_ped(data = data_1, Surv(obs_times, status) ~ ., id = "id",
                cut = seq(0, max(data_1$obs_times), 0.4))

data_1
```

id	obs_times	status
1	1.0	0
2	0.5	1
3	2.0	1

```r
ped_1
```

id	tstart	tend	interval	offset	ped_status
1	0.0	0.4	(0,0.4]	-0.9163	0
1	0.4	0.8	(0.4,0.8]	-0.9163	0
1	0.8	1.2	(0.8,1.2]	-1.6094	0
2	0.0	0.4	(0,0.4]	-0.9163	0
2	0.4	0.8	(0.4,0.8]	-2.3026	1
3	0.0	0.4	(0,0.4]	-0.9163	0
3	0.4	0.8	(0.4,0.8]	-0.9163	0
3	0.8	1.2	(0.8,1.2]	-0.9163	0
3	1.2	1.6	(1.2,1.6]	-0.9163	0
3	1.6	2.0	(1.6,2]	-0.9163	1

The offset reflects information on the observed survival time. Bender et al. (2018a) outline that by re-arranging the Poisson likelihood of the PEM one can derive the following definition:

$$\exp(o_{ij}) = t_{ij} \tag{2.4}$$

where o_{ij} reflects the offset of individual i at time j and t_{ij} to observed failure time. Effectively, this means that the offset o_{ij} is equal to the natural logarithm of interval length of the observation of interest i if there is an interval cut for each event time.

$$o_{ij} = \log(t_{ij} - t_{ij-1}) \tag{2.5}$$

If we make use of custom cuts, it is slightly more complex.

If there was no event for individual i in the respective interval, o_{ij} is equal to the natural logarithm of interval length of the observation of interest – like before.

In case there was an event within the interval but not exactly at the cut points, the offset is computed as the natural logarithm of the current interval length plus the natural logarithm of the proportion of the interval that has passed until the observed event. This means for an interval length of 0.4 and an event at $t_{ij} = 0.5$ (in the second interval) this means:

$$o_{ij} = \log(0.4) + \log((0.5 - 0.4)/(0.8 - 0.4)) = \log(0.4) + \log(0.25) = -2.302585$$

2.2.4.2 Static covariates

Static covariates are very easily incorporated in the PEM format as they only need to be **repeated** for every interval of an `id`.

```r
data_2 <- data.frame(id = 1:3,
                     obs_times = c(1, 0.5, 2),
                     status = c(0, 1, 1),
                     x1 = c(0, 4, 5),
                     x2 = c(1, 1, 2))
ped_2 <- as_ped(data = data_2, Surv(obs_times, status) ~ ., id = "id",
                cut = seq(0, max(data_2$obs_times), 0.4))

data_2
```

id	obs_times	status	x1	x2
1	1.0	0	0	1
2	0.5	1	4	1
3	2.0	1	5	2

```
ped_2
```

id	tstart	tend	interval	offset	ped_status	x1	x2
1	0.0	0.4	(0,0.4]	-0.9163	0	0	1
1	0.4	0.8	(0.4,0.8]	-0.9163	0	0	1
1	0.8	1.2	(0.8,1.2]	-1.6094	0	0	1
2	0.0	0.4	(0,0.4]	-0.9163	0	4	1
2	0.4	0.8	(0.4,0.8]	-2.3026	1	4	1
3	0.0	0.4	(0,0.4]	-0.9163	0	5	2
3	0.4	0.8	(0.4,0.8]	-0.9163	0	5	2
3	0.8	1.2	(0.8,1.2]	-0.9163	0	5	2
3	1.2	1.6	(1.2,1.6]	-0.9163	0	5	2
3	1.6	2.0	(1.6,2]	-0.9163	1	5	2

2.2.4.3 Time-dependent covariates

There are covariates which require more pre-processing, though, namely time-dependent covariates. Time-dependent covariates need to be accounted for by incorporating these times at which variable values change as intervals in the PED format. Typically, longitudinal data comes in an additional separate data set which can be joined by a primary key to the data set without longitudinal data. This relational data model stores the data in the most efficient manner. The data sets below are an example of this. We now assume that `x2` changes over time.

```
data_3 <- data.frame(id = 1:3,
                     obs_times = c(1, 0.5, 2),
                     status = c(0, 1, 1),
                     x1 = c(0, 4, 5))
data_3_long <- data.frame(id = c(1, 1, 2, 2, 2, 2, 3, 3, 3),
                          day = c(0, 1,
                                  0, 0.2, 0.3, 0.4,
                                  0, 1, 2),
                          x2 = c(1, 1,
                                 1, 0, 2, 1,
                                 0, 1, 2))

data_3
```

id	obs_times	status	x1
1	1.0	0	0
2	0.5	1	4
3	2.0	1	5

```r
data_3_long
```

id	day	x2
1	0.0	1
1	1.0	1
2	0.0	1
2	0.2	0
2	0.3	2
2	0.4	1
3	0.0	0
3	1.0	1
3	2.0	2

The new PEM format data has now more intervals and the intervals are not necessarily equidistant anymore, as discrete change moments are explicitly recognised.

```r
ped_3 <- as_ped(
  data    = list(data_3, data_3_long),
  formula = Surv(obs_times, status) ~ . + concurrent(x2, tz_var = "day"),
  id      = "id",
  cut     = seq(0, max(data_3$obs_times), 0.4))

ped_3
```

id	tstart	tend	interval	offset	ped_status	x1	x2
1	0.0	0.2	(0,0.2]	-1.6094	0	0	1
1	0.2	0.3	(0.2,0.3]	-2.3026	0	0	1
1	0.3	0.4	(0.3,0.4]	-2.3026	0	0	1
1	0.4	0.8	(0.4,0.8]	-0.9163	0	0	1
1	0.8	1.0	(0.8,1]	-1.6094	0	0	1
2	0.0	0.2	(0,0.2]	-1.6094	0	4	1
2	0.2	0.3	(0.2,0.3]	-2.3026	0	4	0
2	0.3	0.4	(0.3,0.4]	-2.3026	0	4	2
2	0.4	0.8	(0.4,0.8]	-2.3026	1	4	1
3	0.0	0.2	(0,0.2]	-1.6094	0	5	0
3	0.2	0.3	(0.2,0.3]	-2.3026	0	5	0
3	0.3	0.4	(0.3,0.4]	-2.3026	0	5	0
3	0.4	0.8	(0.4,0.8]	-0.9163	0	5	0
3	0.8	1.0	(0.8,1]	-1.6094	0	5	0
3	1.0	1.2	(1,1.2]	-1.6094	0	5	1
3	1.2	1.6	(1.2,1.6]	-0.9163	0	5	1
3	1.6	2.0	(1.6,2]	-0.9163	1	5	1

2.2.4.4 Cumulative effects covariates

The main objective of Bender et al. (2018b) was the facilitation of the flexible estimation of complex ELRA effects. For an exhaustive explanation of cumulative effects, please refer to the next chapter. For the moment, consider cumulative effects to do the following. Assume that subjects are exposed

– during the follow-up – to some exposure. This exposure can, for example, be radiation that is absorbed by the individuals. Cumulative effects further generalise time-dependent effects by allowing the effect to last over a specific period. In detail, this is achieved by rather viewing the total amount of exposure instead of the newly absorbed one only. This cumulation of exposure is assumed to affect the target altogether. These complex effects can be modeled by encoding the covariates as functional data. A cumulative effect is in principle based on static and / or time-varying covariates. However, to easily model them, they are also preprocessed. Thus technically, this preprocessing is already in some sense a part of the modeling procedure.

```
data_4 <- data.frame(id = 1:3,
                     obs_times = c(3.5, 2.5, 2.25),
                     status = c(0, 1, 1))
data_4_cumu <- data.frame(id = c(1, 1, 1,
                                 2, 2, 2,
                                 3, 3, 3),
                          day = rep(1:3, 3),
                          x1 = c(0, 1, 1,
                                 0, 1, 0,
                                 1, 0, 3))

data_4
```

id	obs_times	status
1	3.50	0
2	2.50	1
3	2.25	1

```
data_4_cumu
```

id	day	x1
1	1	0
1	2	1
1	3	1
2	1	0
2	2	1
2	3	0
3	1	1
3	2	0
3	3	3

From this longitudinal data representation, one can easily construct a PEM representation featuring cumulative effects. To do so, the minimum requirement is to sufficiently supply cumulative effect in the formula indicating the time and the exposure covariate. Technically, in `pammtools` this is achieved by `cumulative(day, x1, tz_var = "day")`. The first two arguments are the covariates on which the cumulative effect will be based. `tz_var = "day"` states which covariate indicates the times at which the covariates were observed.

```
ped_4 <- as_ped(
  data    = list(data_4, data_4_cumu),
  formula = Surv(obs_times, status) ~ . + cumulative(day, x1, tz_var = "day"),
  cut      = seq(0, max(data_4$obs_times), 0.5),
  id = "id")
```

```
ped_4[, c("id", "tend", "ped_status", "day", "x1", "LL")]
```

id	tend	ped_status	day.1	day.2	day.3	x1.1	x1.2	x1.3	LL.1	LL.2	LL.3
1	0.5	0	1	2	3	0	1	1	0	0	0
1	1.0	0	1	2	3	0	1	1	0	0	0
1	1.5	0	1	2	3	0	1	1	1	0	0
1	2.0	0	1	2	3	0	1	1	1	0	0
1	2.5	0	1	2	3	0	1	1	1	1	0
1	3.0	0	1	2	3	0	1	1	1	1	0
1	3.5	0	1	2	3	0	1	1	1	1	1
2	0.5	0	1	2	3	0	1	0	0	0	0
2	1.0	0	1	2	3	0	1	0	0	0	0
2	1.5	0	1	2	3	0	1	0	1	0	0
2	2.0	0	1	2	3	0	1	0	1	0	0
2	2.5	1	1	2	3	0	1	0	1	1	0
3	0.5	0	1	2	3	1	0	3	0	0	0
3	1.0	0	1	2	3	1	0	3	0	0	0
3	1.5	0	1	2	3	1	0	3	1	0	0
3	2.0	0	1	2	3	1	0	3	1	0	0
3	2.5	1	1	2	3	1	0	3	1	1	0

The cumulative effect specified here is assumed to permanently affect the outcome. The effect of the exposure starts to affect the target in the next period after the exposure was absorbed. (This means that there is a by default built-in lag already.) While this special case seems meaningful for some applications, we would like to have a further generalisation. This can be achieved by defining a lag-lead-function which specifies *how* the cumulative effect actually affects the outcome, or in other words, **how cumulative it is**. Lag-lead-functions need to be defined as a function concerning the follow-up time t (`t`) and the time of observation of the time-dependent covariate t_z (`tz`).

Next to the default case, we will present two different lag-lead functions. The second one will be the case where the cumulative effect only affects the outcome in the period of the exposure (i.e. this is an ordinary time-dependent effect). The third one will incorporate a pre-defined lag. However, by defining custom lag-lead functions many more complex configurations are feasible.

First, we define the default lag-lead function from the `cumulative` function and plot it on a 10x10 (follow-up by exposure time) grid. Note that the grid is arbitrarily chosen and will always adapt to the specific data situation.

```
ll_fun1 = function(t, tz) t >= tz
ll1 <- get_laglead(0:10,
                   tz = 1:10,
                   ll_fun = ll_fun1)
gg_laglead(ll1)
```

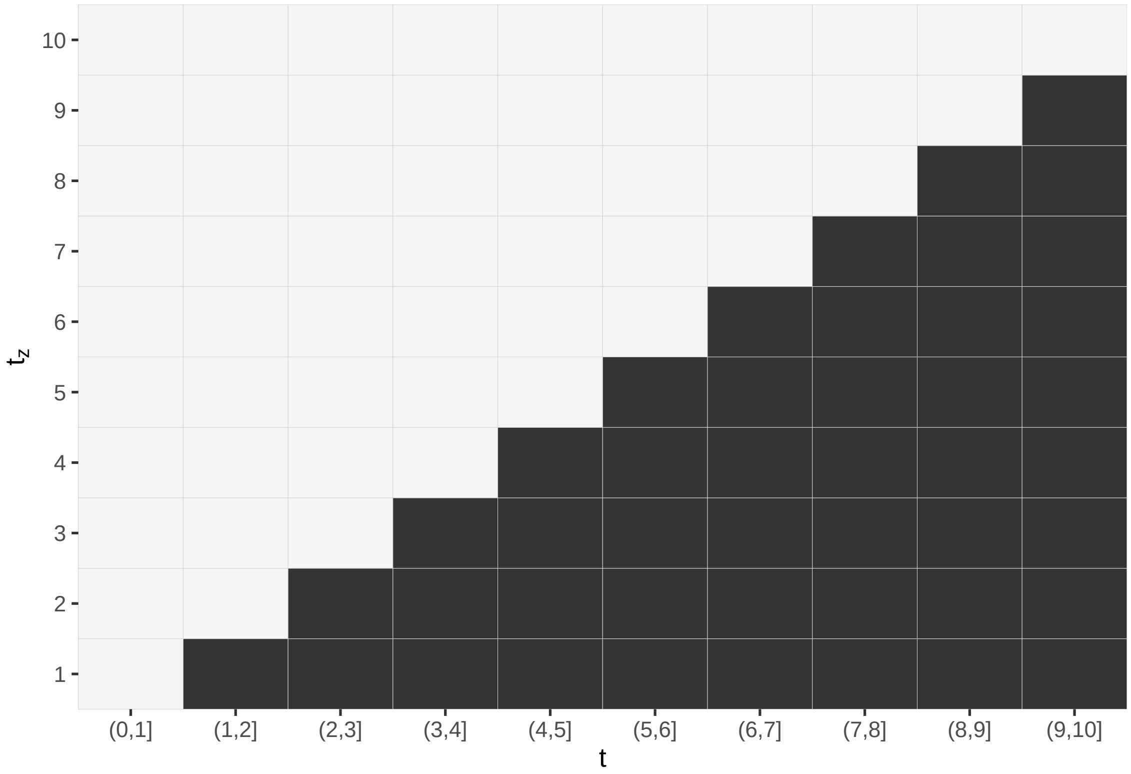

Figure 2.1: Example for a lag-lead window. There is a default lag of 1. After the lag each period affects the outcome until the last period.

The (trivial) case of an ordinary TDC can be constructed by (the `- 1` makes sure that already in the period of exposure there will be an effect):

```r
ll_fun2 = function(t, tz) t == tz - 1
ll2 <- get_laglead(0:10,
                   tz = 1:10,
                   ll_fun = ll_fun2)
gg_laglead(ll2)
```

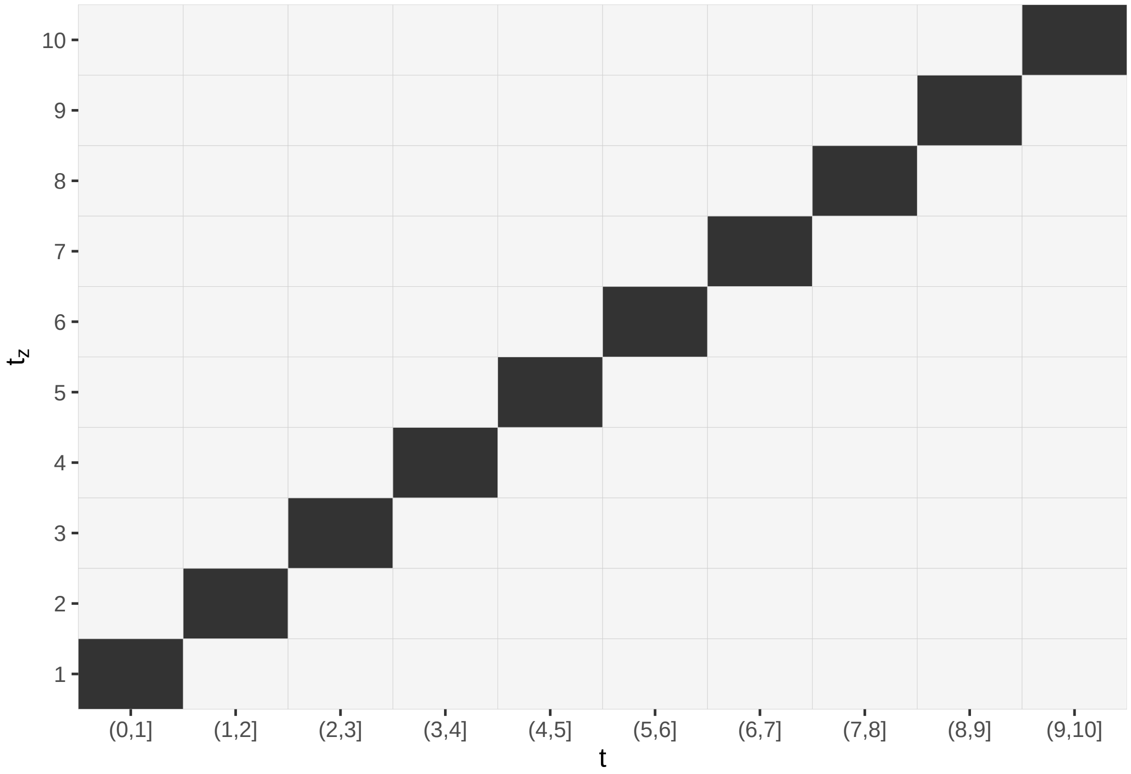

Figure 2.2: Example for a lag-lead window. Each period only affects the outcome for the same period.

A fixed lag window (of 2) can be defined by, e.g.:

```
ll_fun3 = function(t, tz) t >= tz + 2 - 1
ll3 <- get_laglead(0:10,
                   tz = 1:10,
                   ll_fun = ll_fun3)
gg_laglead(ll3)
```

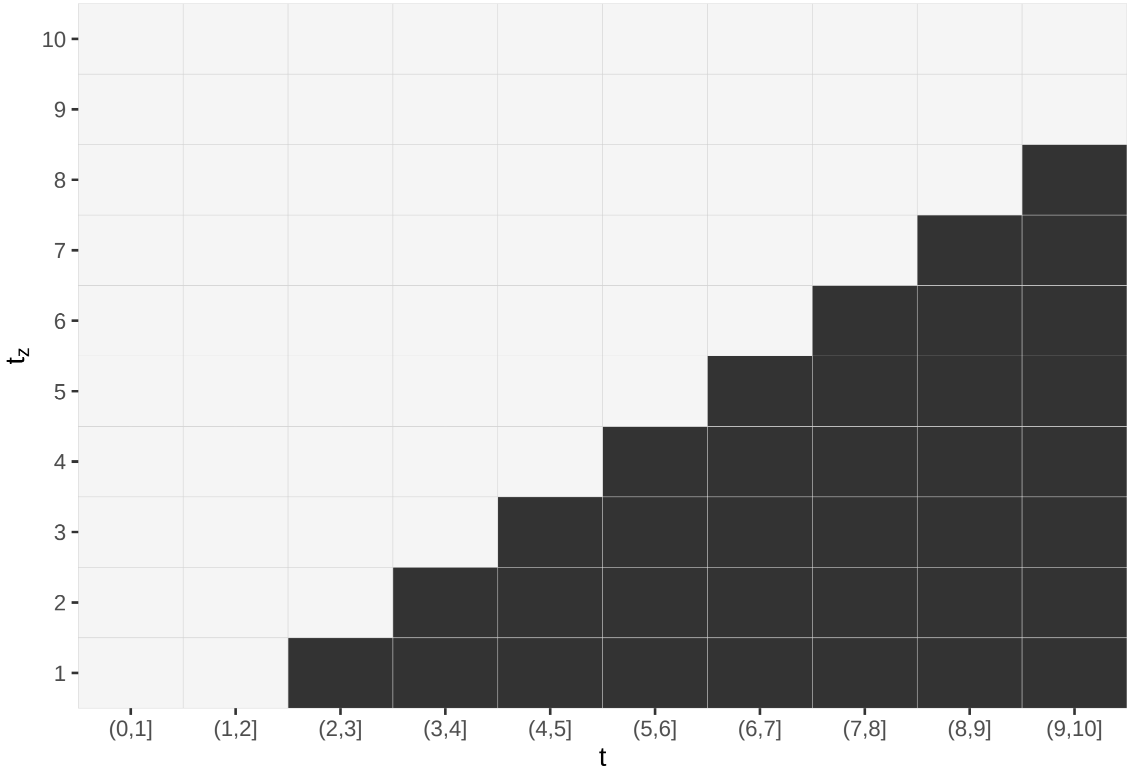

Figure 2.3: Example for a lag-lead window. There is a default lag of 1. After the lag each period affects the outcome for one period.

Using the first lag-lead function, we end up with the following PAM data frame.

```
ped_5 <- as_ped(
  data    = list(data_4, data_4_cumu),
  formula = Surv(obs_times, status) ~ . +
    cumulative(day, x1, tz_var = "day", ll_fun = ll_fun1),
  cut     = seq(0, max(data_4$obs_times), 0.5),
  id = "id")
```

```
ped_5[, c("id", "tend", "ped_status", "day", "x1", "LL")]
```

id	tend	ped_status	day.1	day.2	day.3	x1.1	x1.2	x1.3	LL.1	LL.2	LL.3
1	0.5	0	1	2	3	0	1	1	0	0	0
1	1.0	0	1	2	3	0	1	1	0	0	0
1	1.5	0	1	2	3	0	1	1	1	0	0
1	2.0	0	1	2	3	0	1	1	1	0	0
1	2.5	0	1	2	3	0	1	1	1	1	0
1	3.0	0	1	2	3	0	1	1	1	1	0
1	3.5	0	1	2	3	0	1	1	1	1	1
2	0.5	0	1	2	3	0	1	0	0	0	0
2	1.0	0	1	2	3	0	1	0	0	0	0
2	1.5	0	1	2	3	0	1	0	1	0	0
2	2.0	0	1	2	3	0	1	0	1	0	0
2	2.5	1	1	2	3	0	1	0	1	1	0
3	0.5	0	1	2	3	1	0	3	0	0	0
3	1.0	0	1	2	3	1	0	3	0	0	0
3	1.5	0	1	2	3	1	0	3	1	0	0
3	2.0	0	1	2	3	1	0	3	1	0	0
3	2.5	1	1	2	3	1	0	3	1	1	0

Note that t_z day 0 is taking into account when viewing the following matrix where we can inspect the (raw) cumulative effect for individual 1.

```
mat <- ped_5$LL * ped_5$x1
colnames(mat) <- c("t_z_1", "t_z_2", "t_z_3")
rownames(mat) <- paste("t_", ped_5$tstart, sep = "")
mat[0:7, ]
```

	t_z_1	t_z_2	t_z_3
t_0	0	0	0
t_0.5	0	0	0
t_1	0	0	0
t_1.5	0	0	0
t_2	0	1	0
t_2.5	0	1	0
t_3	0	1	1

where we can see which days or effects are cumulated. To receive the “current” cumulated amount, on needs to determine `rowSums()`.

```
rowSums(mat[0:7, ])
```

```
##   t_0 t_0.5   t_1 t_1.5   t_2 t_2.5   t_3
##     0     0     0     0     1     1     2
```

2.2.5 Model properties and specifications

Using a PAMM one models the log hazard $\log(\lambda_i(t|X))$ for a single individual i.

A very general PAMM looks like the following:

$$\log(\lambda_i(t|X)) = f(x_{i,p}) \tag{2.6}$$

Following Bender and Scheipl (2018) one can specify $f(x_{i,p})$ in a very flexible manner. Basically, one can specify all effects that are possible using `mgcv`. The following table from Bender and Scheipl (2018) depicts many possible effects:

Table 2.1: Possible effect specifications in PAMMs / GAMMs using `mgcv`.

$f(x_{i,p})$	Effect	`mgcv` specification
$\beta_p x_{i,p}$	Linear, time-constant	`... + x + ...`
$f_p(x_{i,p})$	Smooth, nonlinear, time-constant	`... + s(x) + ...`
$\beta_p x_{i,p} + \beta_{p:t} x_{i,p} t$	Linear, linearly time-varying	`... + x + x:t +...`
$f_p(x_{i,p})t$	Smooth, linearly time-varying	`... + s(x, by = t) + ...`
$x_{i,p} f_p(x_{i,p})$	Linear, smoothly time-varying	`... + s(t, by = x) + ...`
$f_p(x_{i,p}, t)$	Smooth, smoothly time-varying	`... + te(x, t) + ...`

Model effects over time can use different time covariates. For example, `tstart` or `tend` can be used. However, other time covariates not directly originating from the follow-up time are possible.

2.2.6 Estimation

In contrast to Cox model survival analysis where only regression coefficients are estimated using (partial) Maximum Likelihood, the baseline hazards are also estimated with likelihood-based inference in the case of piece-wise exponential models.

For piece-wise exponential models we achieve parameter estimates using a Poisson likelihood. A single log likelihood contribution looks like the following (Bender (2018)):

$$\begin{aligned}\ell_i(\beta) &= \log\left(\prod_{j=1}^{j(i)} f\left(\delta_{ij}\right)\right) = \sum_{j=1}^{j(i)} \delta_{ij} \log\left(\mu_{ij}\right) - \mu_{ij} \\ &= \sum_{j=1}^{j(i)} \delta_{ij} \log\left(\lambda_{ij}\right) + \delta_{ij} \log\left(t_{ij}\right) - \lambda_{ij} t_{ij}\end{aligned} \tag{2.7}$$

where δ_{ij} is the event indicator for individual i in the current interval j. The coefficients β are plugged into the likelihood contribution; then, one optimises for β and receives (semi-)parametric estimates. All parameter estimates follow a normal distribution. This property is used for example when deriving confidence intervals later on.

Typically, piece-wise exponential models require more data to be (reliably) estimated than equivalent Cox models as the (semi-)parametric estimation of the baseline leaves fewer degrees of freedom.

Covariate effects can be estimated in the same way as in an ordinary GAM. Their interpretation is the same as in the Cox model. Estimating the baseline hazard is less intuitive and is presented briefly using the example from the `pammtools` vignette.

2.2.7 Estimating the baseline hazard

The baseline hazard is estimated parametrically when dealing with a PEM. Then, each interval has its own parameter which is not different w.r.t. the estimation compared to regression coefficients.

We illustrate the modeling of a parametric baseline by a data example following the `pammtools` vignette.

```
data("veteran", package = "survival")
```

We use the `veteran` data. The data set features information on a lung cancer study where there is a randomised treatment and control group. All patients in the study suffer from lung cancer. The survival time and the survival status (`status`) are tracked for all patients (`time`). Next to information on the treatment status (`trt`) the data contains information on the type of lung cancer (`celltype`), the age of the patients (`age`), and more. For details, consider the Appendix or the documentation of the data set.

We want to estimate the baseline hazard of veterans who took part in the RCT – no matter which treatment has been assigned and neglecting all other covariates. Following the vignette, we filter the data. The last filter – the removal of duplicated event times (or ties) – is necessary to achieve the exact equivalence of the PEM with the Cox model.

```
veteran_new <- veteran %>%
  mutate(
    trt   = 1 * (trt == 2),
    prior = 1 * (prior != 10)) %>%
  filter(time < 400)
ped <- veteran_new %>%
  as_ped(Surv(time, status) ~ ., id = "id")
```

We model a PEM as an ordinary `glm` with the `family` argument set to `poisson()`. We have to explicitly consider the `offset` to account for the lengths of intervals. The only parameter to be modeled in the `glm` is the interval. We end up with one estimate for each `level` of the `factor` variable `interval` from `ped`.

```
pem <- glm(ped_status ~ interval, data = ped, offset = offset, family = poisson())
```

When calling the summary, we will receive a very long output with non-meaningfully interpretable estimates for single intervals. This is why we do not show the model summary here.

Typically, we would suggest supplying custom cut points which however match the data situation well. Consider Bender (2018) for an empirical discussion on the placing of cut points.

We can use the model to predict hazards, or even more informative cumulative hazards and survival functions. To predict these in `R` we need to supply to the convenience function `add_cumu_hazard()` what we actually want to predict for. For the PEM just estimated (`pem`) this will be all the non-overlapping intervals (`interval`) used for the modeling in the `glm`. `add_cumu_hazard()` calls the `predict.glm()` method and derives hazards for each interval. These hazards are transformed into cumulative hazards straightforwardly by adding them up. Another convenience function, `int_info()` constructs a data frame that we can use as input for `add_cumu_hazard()`, supplying all intervals.

```r
int_df <- int_info(ped)
head(int_df)
```

tstart	tend	intlen	intmid	interval
0	1	1	0.5	(0,1]
1	2	1	1.5	(1,2]
2	3	1	2.5	(2,3]
3	4	1	3.5	(3,4]
4	7	3	5.5	(4,7]
7	8	1	7.5	(7,8]

```r
int_df <- int_df %>% add_cumu_hazard(pem, ci = FALSE)
head(int_df)
```

tstart	tend	intlen	intmid	interval	cumu_hazard
0	1	1	0.5	(0,1]	0.0153
1	2	1	1.5	(1,2]	0.0230
2	3	1	2.5	(2,3]	0.0308
3	4	1	3.5	(3,4]	0.0387
4	7	3	5.5	(4,7]	0.0625
7	8	1	7.5	(7,8]	0.0950

We can compare this parametric estimate to the non-parametric estimate retrieved by the Nelson-Aalen estimate.

```r
library(survival)
base_df <- basehaz(coxph(Surv(time, status) ~ 1, data = veteran_new)) %>%
  rename(nelson_aalen = hazard)
int_df <- int_df %>% add_column(Model = "PEM")
int_df <- rbind(
  int_df,
  int_df %>% left_join(base_df, by = c("tend" = "time")) %>%
    select("tstart", "tend", "intlen", "intmid", "interval", "nelson_aalen") %>%
    rename(cumu_hazard = nelson_aalen) %>% add_column(Model = "Nelson-Aalen"))
```

When we plot both estimated curves next to one another, we can see the equivalence.

```r
library(ggplot2)
ggplot(int_df, aes(x = tend)) +
  geom_step(aes(y = cumu_hazard, col = Model, linetype = Model), size = 1.05) +
  theme(legend.position = "bottom") +
  ylab(expression(hat(Lambda)(t))) + xlab("t") +
  ggtitle(paste("Comparison of cumulative hazards estimated by Cox-PH",
                "vs. PEM \n Analyis of mortality in the veteran data"))
```

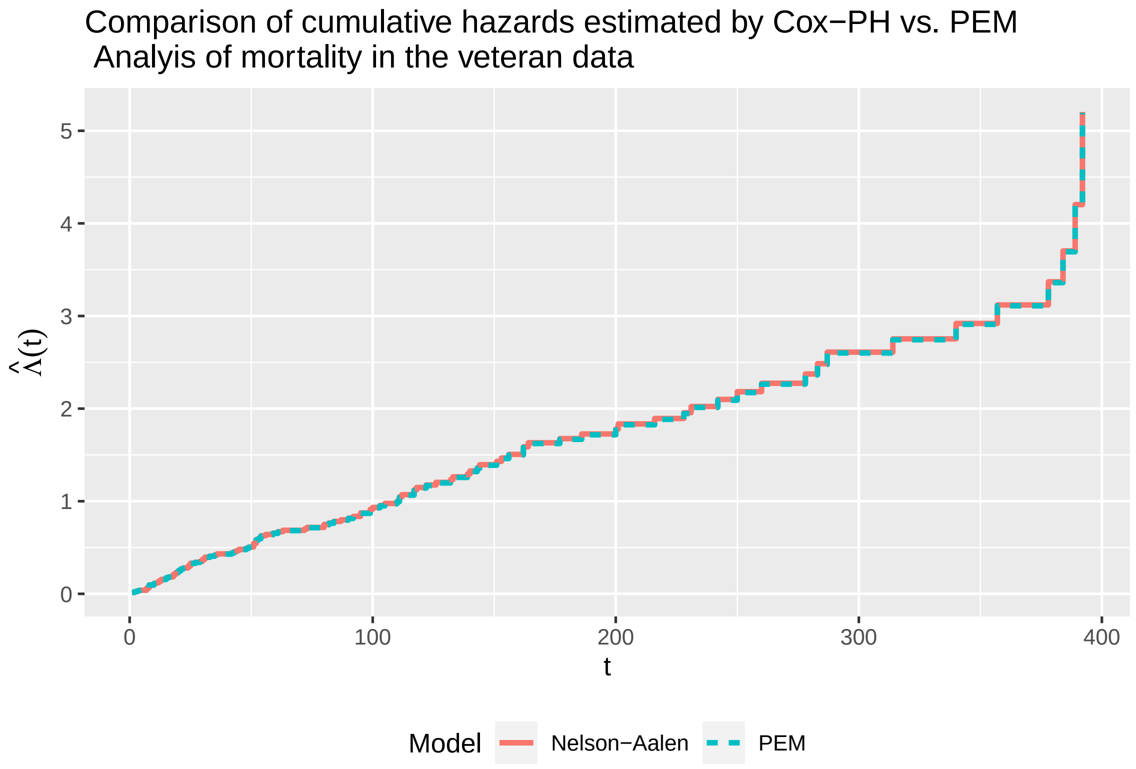

Figure 2.4: Comparison of cumulative hazards rates between parametric PEM estimates and non-parametric Nelson-Aalen estimates

When modeling a smooth baseline, though, we technically model spline smooths in a `gam`. A smooth baseline is very desirable for two reasons. First, a smooth baseline better reflects the natural risk process for most risk processes. Second, the selection of the cut points is no crucial parameter. Empirically, we find that PEMs can produce very high standard errors if the number of intervals is large. PAMs – featuring a smooth baseline – typically use significantly less effective degrees of freedom. Hence, they produce more reasonable standard errors.

Having used `pammtools` for the preprocessing one can fit a smooth baseline using the `gam()` function from the `mgcv` package (note that also other packages can be used). We only use the `mgcv` package to estimate PAMs/PAMMs/PEMMs and the `stats` package for PEMs. As a PEM can be also formulated as a PAM (with no smooth terms), we focus on the `mgcv` package. `mgcv` uses an approach (by default) called generalised cross-validation (`mgcv` basically performs leave-one-out cross-validation to select parameters) to minimise the penalised likelihood (Clark (2019)):

$$l_p(\beta) = l(\beta) - \alpha B^T S B \tag{2.8}$$

where α is a penalisation parameter. S is a "penalty matrix of known coefficients" (Clark (2019)). This means that S incorporates already the structure of the effect. B represents the basis functions implied by the `gam`.

The smooth terms are optimised subject to a penalty while the non-smooth terms are optimised like in a GLM. For numerical optimisation `mgcv` offers different optimisers. By default, a Newton-Raphson optimiser which needs the first two derivates of the penalised likelihood. (Note that GAMs can, however, be fit with a variety of packages.)

The `s()` function defines a smoothing spline. For a discussion of smoothing splines please refer to Hastie (2017).

```
library(mgcv)
pam <- gam(ped_status ~ s(tend), data = ped, offset = offset, family = poisson())
```

Instead of plotting the cumulative hazards, one could also use the convenience function `add_surv_prob()` to obtain survival function estimates.

```
int_df <- int_info(ped)
int_df <- int_df %>%
  add_surv_prob(pam, ci = TRUE)
```

Now we also want to predict the confidence intervals. The survival function is a non-linear transformation of the cumulative hazard function.

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

Confidence intervals can be computed in three different manners:

a) Simulation from the ex posteriori normally distributed coefficients.

b) Delta Method.

c) Simple (linear) transformation of the linear predictor.

Option c) is the default in the `pammtools` package.

For an empirical comparison of the three methods, refer to Bender (2018).

```
ggplot(int_df, aes(x = tend)) +
  geom_line(aes(y = surv_prob), col = "black") +
  geom_line(aes(y = surv_lower), col = "red", linetype = "dashed") +
  geom_line(aes(y = surv_upper), col = "red", linetype = "dashed") +
  ylab(expression(hat(S)(t))) + xlab("t") +
  ggtitle(paste("Estimated survival function using a PAM",
                "\n Analyis of mortality in the veteran data"))
```

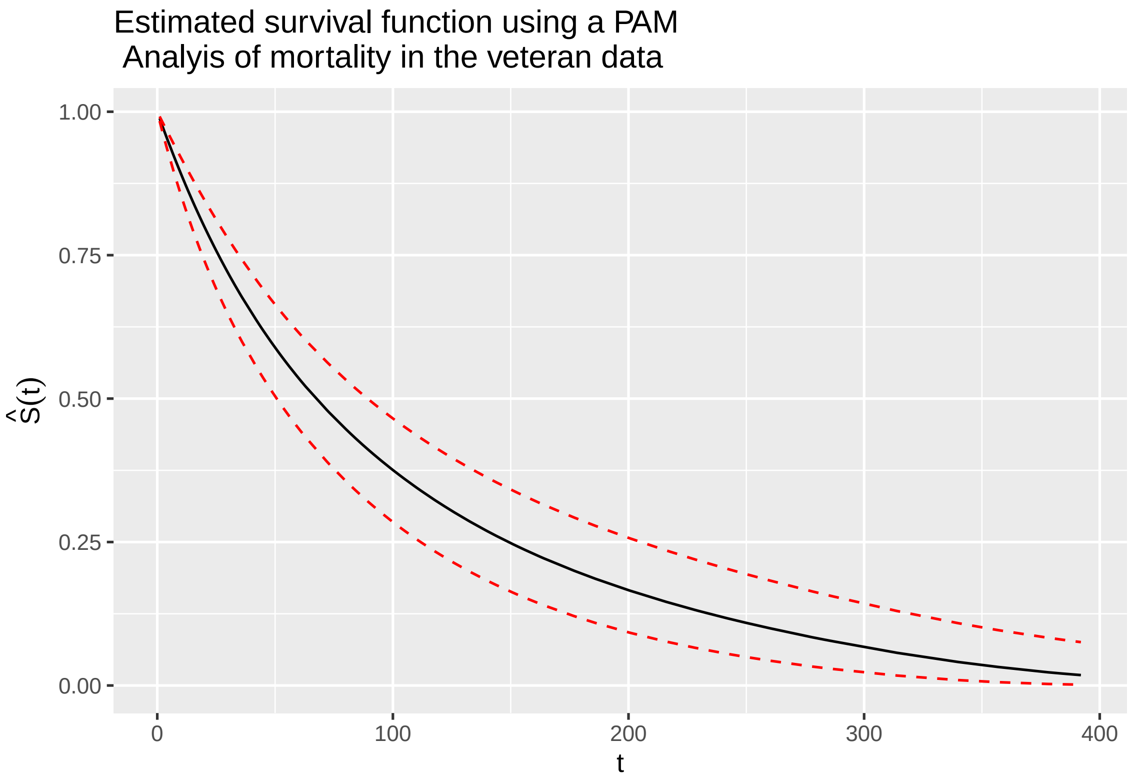

Figure 2.5: Comparison of cumulative hazards rates between parametric PEM estimates and non-parametric Nelson-Aalen estimates

How to deal with covariate effects in `pammtools` can be reviewed here and is left out at this point to avoid redundancies.

Chapter 3

Penalized estimation of complex, non-linear exposure-lag-response associations

This chapter is a detailed summary of Bender et al. (2018b). While Bender et al. (2018b) is a methodological paper, Hartl et al. (2019) feature the same analysis however with a focus on the clinical insights. We will mostly refer to Bender et al. (2018b) throughout this chapter.

The authors analyse complex exposure-lag-response associations of intensive care units' artificial nutrition diet schemes on survival in hospital. Bender et al. (2018b) outline four reasons what makes it challenging to model the association of the intake of artificial nutrition and survival times. The amounts of nutrition vary over time (1) with potentially time-varying effects on the hazard rates (2). These effects may have a lasting impact on the hazard rates (3). Additionally, there is a potential lag of these associations (4). They find that high-caloric nutrition is preferable to increase the survival times of the patients.

3.1 Data

Bender et al. (2018b) analyse data from 457 intensive care units (ICU) in 39 countries. The data source has first been used by Heyland et al. (2011) who originally used them for the investigation of the identical research question as Bender et al. (2018b).

The data used in their analysis is a subset of a large international point prevalence survey of nutrition practice in ICUs ICUs all around the world participated over several years in the study by observing 20 distinct patients who were consecutively intubated, over the age of 18 and mechanically ventilated within the first 48 h in the ICU. The patients had to remain at least 72 h in the ICU to be considered in the study. For each patient, relevant information on the patient status (sex, weight, height, etc.), the health status (Apache II score, admission category, diagnosis category, etc.) have been collected on admission. Soon after admission nutrition goals (protein and calories) have been determined by a dietitian or physician. During the time in the ICU, the nutrition protocols are tracked daily for maximally 12 days. The study distinguishes between oral intake and both, enteral and parenteral nutrition. Further details are explained in Hartl et al. (2019).

In total, the data features 9661 critical care patients with maximally 12 patient days for each. Before Bender et al. (2018b) a subset of the data has already been used by Heyland (2011). The final – ready to use – data is provided by Bender et al. (2018b) online.

We refrain from an exhaustive description of the data. This is for two reasons. First, in chapter 7 there will be a detailed summary of the updated data (which completely includes this data). We aim to avoid redundancies Second, the descriptive analyses of Bender et al. (2018b) can be retrieved from their publications straight away.

3.2 Modeling

The general (abstract) research question of Bender et al. (2018b) is the link between artificial nutrition and patient survival. As survival times (and not only status) are observed, the survival time is used as the dependent variable in our regression set up. More precisely, Bender et al. (2018b) aim to model the 30-days-survival (i.e. everything beyond the scope of 30 days is neglected). We can identify two sets of covariates in their analysis:

1) Confounders

2) Nutritional covariates of special interest: to be precise caloric adequacy.

Dealing with the statistical model – in a machine learning fashion – as a black box for the moment, Bender et al. (2018b) aim to model the following:

$$S_i = f(X_i) \tag{3.1}$$

with

$$\begin{gathered} X_i = (x_{i,year}, x_{i,diag}, x_{i,admission}, x_{i,gender}, \\ x_{i,MV}, x_{i,propofol}, x_{i,oral}, x_{i,parenteral}, \\ x_{i,Apache}, x_{i,age}, x_{i,BMI}, x_{i,nutrition}, x_{i,ICU}) \end{gathered}$$

where:

- $x_{i,year}$ is a categorical variable indicating the year of admission of for patient i.
- $x_{i,diag}$ refers to the diagnosis category.
- $x_{i,admission}$ refers to the admission category.
- $x_{i,gender}$ refers to patient's sex (categorical).
- $x_{i,MV}$ refers to the frequency of mechanical ventilation between days two and four.
- $x_{i,propofol}$ refers to the frequency of propofol administering between days two and four.
- $x_{i,oral}$ refers to the frequency of oral intake of food between days two and four.
- $x_{i,parenteral}$ refers to the frequency of parenteral artificial nutrition between days two and four.
- $x_{i,Apache}$ refers to the Apache II score as reported at admission.

- $x_{i,age}$ refers to the patient's age (at admission).
- $x_{i,BMI}$ refers to the patient's BMI (at admission).
- $x_{i,nutrition}$ refers to the nutritional intake of the patient.
- $x_{i,ICU}$ refers to a categorical variable indicating the ICU.

Both, the nutrition feature and some of the confounders are time-varying. The time-varying confounders are, however, summarised as static values that represent these confounders in the early days of the study. Some of the effects are assumed to be (possibly) non-linear or smooth. Furthermore, the effect for the nutritional intake is assumed to be very complex – an exposure-lag-response association. Additionally, random effects for the different study centers / ICUs need to be facilitated. How these (complex) effects are modeled in general is examined in the next three subsections with a strong focus on exposure-lag-response associations (ELRA).

3.2.1 Exposure-lag-response associations

The most interesting effect to be studied is the effect of caloric intake via artificial nutrition. Nevertheless, the time-dependent indicator covariate has some specific modeling implications.

First, Bender et al. (2018b) assume that the effect of artificial nutrition is not immediate but lagged. Second, Bender et al. (2018b) emphasise that the diet as a whole is more meaningful to be analysed than daily caloric intake. The diet as a whole can be represented by the accumulation of caloric intake over multiple days.

Bender et al. (2018b) propose an exposure-lag-response association (Gasparrini (2014)) (ELRA) to facilitate this kind of effect. In many biostatistical applications, this ELRA assumption is presumed and seems well reasoned. ELRA manages to capture dependencies of the target on both, intensity and timing of exposures.

ELRA is a special functional association (refer to Kokoszka and Reimherr (2017) for an exhaustive introduction to functional data analysis). Functional data applies to very different contexts. One of these is the investigation of time-varying covariates. Functional data approaches often incorporate the time structure into the modeled effects.

3.2.1.1 Derivation

This section is occupied with a general derivation of the type of effect. How this effect is applied in Bender et al. (2018b), is presented in the subsequent section.

Gasparrini (2014) define the association within the function $s(x, t)$ where x is the exposure history and t the evaluation time. The function as a whole can be treated as a covariate in a regression model. However, the data format of this functional data needs to be accounted for.

Gasparrini (2014) derive ELRA from exposure response relationships. They define the functional effect for a time interval of positive length $[t_0, T]$ (t_0 is the first relevant exposure time and T is the last one) as:

$$s(x,t) = \int_{t_0}^{T} x_u w(t-u) du \tag{3.2}$$

where x_u refers to the exposure *intensity* at $t = u$ and $w()$ refers to a weighting function. Typically, this function is estimated using data.

To achieve an expression in terms of lags, Gasparrini (2014) suggest the following reformulation

$$s(x,t) = \int_{l_0}^{L} x_{t-l} w(l) dl \tag{3.3}$$

Using the following identity:

$$L - l_0 = T - t_0 \tag{3.4}$$

with $[l_0, L]$ representing lag period w.r.t. the outcome. Hence, $L - l_0$ is the lag period in which the exposure (presumably) affects the outcome. $w()$ being defined in terms of lags l now refers to the lag-response-function which determines the lag-response curve $s(x,t)$. Typically, dealing with data $s(x,t)$ can be approximated by:

$$s(x,t) \approx \sum_{l=l_0}^{L} x_{t-l} w(l) \tag{3.5}$$

In order to derive the ELRA expression Gasparrini (2014) add two extensions to the previously shown association. The first one is the generalisation of the predictor replacing x_{t-l} by $f(x_{t-l}))$ to account for potential non-linearity of the exposure-response curve $s(x,t)$ through x.

$$s(x,t) = \int_{l_0}^{L} f(x_{t-l}) w(l) dl \tag{3.6}$$

However, the expression $f(x)w(l)$ with the two separate functions $f(x)$ and $w(l)$ is problematic. Implicitly this formulation requires $f(x)$ and $w(l)$ to be independent. Gasparrini (2014) describe that this means "that the exposure-response shape is the same at each lag l, and vice versa that the lag structure is the same at each value of x." Furthermore, the expression is problematic as it cannot be represented (straightforwardly) as a linear combination of basis variables (Gasparrini (2014)). This means that one may need to move classical GAMMs to model these associations being non-linear in the coefficients. As a solution, Gasparrini (2014) present the bivariate **exposure-lag-response function** $fw(x,l)$ (ELR function).

$$s(x,t) = \int_{l_0}^{L} fw(x_{t-l}, l) dl \tag{3.7}$$

As usual, dealing with a discrete-time scale all integrals become sums and the continuous-time t can be replaced by an indicator for time intervals j

Being bivariate, this function loosens the independence proposition from earlier and allows to model **analoga** of $f(x)$, the exposure-response curve, and $w(l)$, the lag-response curve which form the so-called, exposure-lag-response surface together. $fw(x, l)$ can be represented by a tensor product making it applicable to GAMMs.

3.2.1.2 ELRA in the analysis

While the previous section was very general in deriving ELRA, this section will be aligned with Bender et al. (2018b) and outline how they use ELRA in their analysis. For this purpose, this section will use different notation and terminology which is more applicable to the analysis. ELRA can be used for both, static and time-dependent covariates (TDC). Nevertheless, the case of static covariates (i.e. constant exposure intensity) does not require very complex effect estimation. Hence, we assume the more general case of TDC.

Bender et al. (2018b) further generalise ELRA by replacing x in the previous formula by the exposure history Z. This makes the modelling of the intended cumulative effects straight forward. Hence, for one individual, the cumulative ELRA is abstractly denoted by $s(Z_i(t), t)$. Bender et al. (2018b) choose an even more general representation of ELRA by introducing two different time scales: t, the *time* at risk and t_e the **exposure** *time*. t_e is the time at which a covariate is observed, while t is the time at which the dependent variable is evaluated. The exposure history $Z_i(t)$ is defined by the exposure intensities $z_i(t_e)$ at exposure time t_e. For classical longitudinal data $z_i(t_e)$ would be simply described as the value of the covariate at t_e. The lag-response function $w(l)$ can also be parametrised in such manner that one effectively defines a time window $T_e(t)$. The dependent variable y_{it} at time t is affected by the exposure intensity $z_i(t)$ indirectly via the exposure history $Z_i(t)$ over the window $T_e(t)$. Or:

$$Z_i(t) := \{z_i(t_e) \in T_e(t)\} \tag{3.8}$$

ELRA as proposed by Gasparrini (2014) look like the following if additionally distinguishing between t and t_e:

$$s(z_i(t_e), t) = f(t_e, t, z_i(t_e)) \tag{3.9}$$

However, Bender et al. (2018b) simplify this trivariate function to:

$$s(z_i(t_e), t) = f(t_e, t)w_{it}(t_e) \tag{3.10}$$

with w_{it} being parametrised so that is effectively creates a time window:

$$w_{it}(t_e) = \begin{cases} z_i(t_e) & \text{if } t_e \in T_e(t) \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

This simplification is allowed if only linear intensity is investigated. Nevertheless, the components not being independent as outlined by Gasparrini (2014) comes at its costs. The simplification causes that the ELRA can be non-linear over the time t and the exposure time t_e but not regarding the exposure intensity $z_i(t)$ (The exposure intensity only enters linearly.). However, Bender et al.

(2018b) mention that they do not need the more general representation in the resulting analysis and point to Wood (2017) who directly model the three-variate function $f(t_e, t, z_i(t_e))$

The cumulative effect is thus represented by:

$$s(Z_i(t_e), t) = \int_{T_e(t)} f(t_e, t) w_{it}(t_e) du = \int_{T_e(t)} f(t_e, t) z_i(t_e) du \tag{3.12}$$

Again, in practice, the integrals can be approximated by sums.

Following the derivation of Gasparrini (2014) the bivariate function $f(t_e, t)$ can be represented using tensor products:

$$f(t_e, t) = \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_{mk} B_m(t_e) B_k(t) \tag{3.13}$$

which create tensor product splines with $B_m(t_c)$ and $B_k(t)$ referring to the basis functions for the respective time periods. γ_{mk} represents the spline coefficients.

The effect window enters the effect via $w_{it}(t_e)$. This window is described by the lag-lead-window which has been presented in the previous section.

3.2.1.3 Estimation & inference

The lag-lead-window or which directly determines function $w_{it}(t_e)$ is not estimated itself and enters the effect straight away as an ordinary covariate. Hence we need to estimate optimal γ_{mk} from

$$s(t_e, t) = \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_{mk} B_m(t_e) B_k(t) w_{it}(t_e) \tag{3.14}$$

This effect is a bivariate tensor product smooth which interacts with the function resulting from the lag-lead window.

The effect is estimated as usually for tensor product smooths in `mgcv` as described in the previous chapter. The resulting coefficients γ_{mk} are normally distributed and allow the down-stream inference with the resulting effect:

$$\hat{s}(t_e, t) = X_Z \hat{\gamma} \tag{3.15}$$

In a `mgcv` formula this sort of effect can be described by `te(t_e, t, by = w)`.

Bender (2018) points to Marra and Wood (2011) for CI estimation for smooth terms and apply their formulation to ELRA.

Inference for cumulative effects is in principle based on the normally distributed single coefficients of the distribution.

The standard errors for the cumulative effect is described by:

$$\hat{SE}(\hat{s}) = \sqrt{diag(X_Z V_{\hat{\gamma}} X_Z^T)} \tag{3.16}$$

γ is the vector of the basis coefficients from all $\gamma_{m,k}$ (with $\hat{\gamma}$ referring to the estimated coefficients). $V_{\hat{\gamma}}$ is the empirical Bayesian covariance matrix of $\hat{\gamma}$. X_Z is a J by $M * K$ matrix with J being the number of intervals in the PEM representation of the follow-up and $M * K$ is the length of the basis coefficient vector. This means that Z is the design matrix of the exposure history associated with the cumulative effect.

With the distributional assumption of the estimates one can easy derive confidence intervals by:

$$\hat{s} \pm z_{1-\alpha/2} \hat{SE}(\hat{s}) \tag{3.17}$$

As $\hat{SE}(\hat{s})$ is a vector of length J, there is an uncertainty estimate for each interval; continuous confidence intervals can be derived by e.g. interpolation.

3.2.1.4 Specification

Bender et al. (2018b) make use of ELRA to measure the effect of artificial nutrition on hospital mortality. The nutrition covariate has been converted to a categorical feature indicating for each day which of three diet schemes (C I, C II, C III):

- C I: less or equal than 30% of prescribed calories.
- C II: between 30% (more than) and 70% (less or equal) of prescribed calories **or** less or equal than 30% with additional oral intake.
- C III: more than 70% of prescribed calories **or** between 30% and 70% with additional oral intake.

Base on this structure, Bender et al. (2018b) construct the different cumulative effects, namely $g_{C_{II}}(t_e, t)$ and $g_{C_{III}}(t_e, t)$. They assess the current exposure of each diet of day t_e on day t via category II diet ($g_{C_{II}}$) or category III diet ($g_{C_{III}}$) covariates. The cumulative exposure is created by the summation, e.g. $\sum_{t_e} g_{c_{II}}(t_e, t)$. Exposures are mutually exclusive so that – like for categorical data – only two effects are identifiable. The third one is characterised by the absence of the other two.

3.2.2 Time-varying effects

Bender et al. (2018b) aim to model TDC very flexibly. They use two different possible effect specifications (see: Bender (2018)):

1) Linear, linearly time-varying effect.

$$\beta_p x_{i,p} + \beta_{p:t}(x_{i,p} t)$$

2) Smooth, smoothly time-varying effect.

$$f(x_{i,p})t$$

All kinds of effects are readily implemented into `mgcv`. The mapping to `mgcv` formulas can be obtained from the previous chapter.

For the Apache II score, Bender et al. (2018b) intend to model a linear, linearly time-varying effect. $\beta_{Apache} * x_{i,Apache} + \beta_{Apache:t} * (x_{i,Apache} * t)$ For age, Bender et al. (2018b) use a smooth, linearly time-varying effect: $f_{age}(x_{i,age})t$ The BMI goes into the model as a smooth, linearly time-varying effect:: $f_{BMI}(x_{i,BMI})t$

3.2.3 Random and fixed effects

Bender et al. (2018b) aim to control for the different ICUs as there may be a great amount of heterogeneity among these. Two prominent ways to do so is by random effects of fixed effects. Fixed effects can be formulated as ordinary covariates in the regression context. However, random effects require special treatment.

Modeling random effects in the Cox PH model relies on the extension of the model to the Hierarchical Cox model (Sargent (1998)). Still, the literature of modeling these kinds of models is not as exhaustive as the mixed models literature using GAMMs.

Random effects for `x` are supplied to `mgcv` via `s(x, bs = "re")`.

Bender et al. (2018b) include a random intercept of the ICUs. b_{l_i} is this random intercept for the different ICUs (as contained in $x_{i,ICU}$)

3.2.4 Model set up

These diverse associations are a major challenge for classical survival models (especially the complex ELRA). The extended Cox model offers has been increasingly extended, so that many complex effects are also facilitated. For example, Danieli and Abrahamowicz (2019) incorporate ELRA into the Cox framework.

However, within the GAMM literature there exist many different effects; Formulating these effects in a PAMM is a no-brainer compared to a cox model formulation. PAMMs make the composed problem of modeling in the Cox environment to two separate ones. Modeling the effects is now a GAMM problem.

3.2.4.1 General set up

As usual in time-to-event analysis the target y_i is the log hazard $\log(\lambda_i(t|X)$ They specify the model in the following manner:

$$
\begin{aligned}
\log(\lambda_i(t|X)) = \\
&\lambda_0(t) + \beta_{year} * x_{i,year} + \beta_{diag} * x_{i,diag} + \\
&\beta_{admission} * x_{i,admission} + \beta_{gender} * x_{i,gender} + \beta_{MV} * x_{i,MV} + \\
&\beta_{propofol} * x_{i,propofol} + \beta_{oral} * x_{i,oral} + \beta_{parenteral} * x_{i,parenteral} \\
&\beta_{Apache} * x_{i,Apache} + \beta_{Apache:t} * (x_{i,Apache} * t) + \\
&f_{age}(x_{i,age})t + \\
&f_{BMI}(x_{i,BMI})t + \\
&\sum_{t_e} g_{c_{II}}(t_e, t) + \sum_{t_e} g_{c_{III}}(t_e, t) + \\
&b_{l_i}
\end{aligned}
\tag{3.18}
$$

where:

- $\lambda_i(t|X)$ is the log hazard of failing (i.e. to die) in hospital w.r.t. to the 30-day survival.
- $\lambda_0(t)$ is the (log) baseline hazard.

All β refer to the standard regression coefficient where the index outlines the relationship to the associated covariate.

As outlined in the previous chapters, complex effects rely on appropriate preprocessing. Especially, cumulative effects need extensive preprocessing. Bender et al. (2018b) has been published before the release of Bender and Scheipl (2018). Hence, the original approach does not make use of the package. However, here we outline how to make use of the `pammtools` package to achieve the intended data set.

First of all, one needs to specify the intended lag-leag window. While Bender et al. (2018b) use different lag-lead windows (the simplest is just a lag of four days with a permanent effect), the main specification is described by a four-day lag which persists over a flexible time window which equals twice the amount of study days. That means the effect of study day 6 persists 18 days.

```r
library(pammtools)
ll_fun = function(t, tz) { t >= tz + 4 & t <= tz * 3 + 12  }
ll <- get_laglead(0:30,
                  tz = 1:11,
                  ll_fun = ll_fun)
gg_laglead(ll) + theme(text = element_text(size = 12),
                       axis.text.x = element_text(angle = 90, hjust = 1))
```

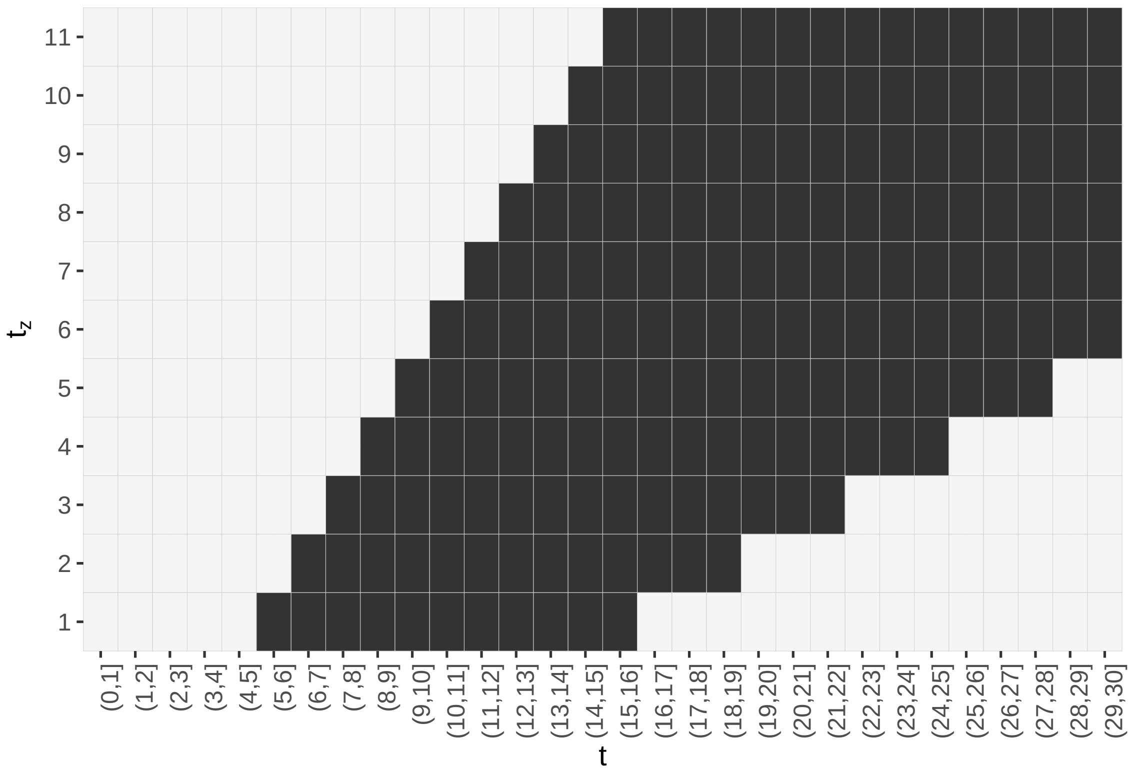

Figure 3.1: Lag-lead window specification in the analysis. The lag-lead window has a minimal lag of 4 days and is dynamically increasing.

Before preprocessing the data via pammtools, we reduce the merged data set to only time-varying factors. This data set is called daily. Furthermore, confounders for parenteral nutrition, mechanical ventilation, and oral intake need to be constructed. Note that we do not perform any of these steps here and only show how to convert this data set to a ped data frame.

```r
ped <- as_ped(
  data    = list(patient, daily),
  formula = Surv(Survdays, PatientDied) ~ Year + DiagID2 + AdmCatID + Gender +
    ApacheIIScore + BMI + CombinedicuID + PN2to4 + MV2to4 + OralIntake2to4,
  cumulative(Survdays, Study_Day, calCat2,
             tz_var = "Study_Day", ll_fun = ll_fun) +
    cumulative(Survdays, Study_Day, calCat3,
               tz_var = "Study_Day", ll_fun = ll_fun),
  cut     = 0:30,
  id = "CombinedID")
```

The final formula which is used for the gam with Poisson likelihood is:

```r
formula = as.formula('ped_status ~ s(int_mid, bs = "ps") +
  ApacheIIScore + ApacheIIScore:int_mid +
  s(Age, by = int_mid, bs = "ps") +
  s(BMI, by = int_mid, bs = "ps") + DiagID2 + AdmCatID + Gender +
  inMV2_4 + Propofol2_4 + OralIntake2_4 + PN2_4 + s(CombinedicuID, bs = "re") +
  te(event_Study_Day_mat, Study_Day, by = I(LL_Study_Day * calCat2_Study_Day),
  bs = "ps", m = list(c(2, 1), c(2, 1)), id = "cal") +
  te(event_Study_Day_mat, Study_Day, by = I(LL_Study_Day * calCat3_Study_Day),
  bs = "ps", m = list(c(2, 1), c(2, 1)), id = "cal")')
```

Whenever `t` is used explicitly, the midpoint of each `ped` interval (computed via `tend - tstart`) is used in this analysis. `I(LL_Study_Day * calCat2_Study_Day)` reconstructs the $w()$ function described earlier where the exposure is mapped with the lag-lead function. `m = list(c(2, 1), c(2, 1))` refers to the order of the spline used and its penalty. `id = cal` for both `te` effects makes sure that the same penalisation is applied for both effects.

3.3 Statistical results

The primary interest of the study was the complex association of nutritional intake and hospital mortality. Hence, both, this thesis and Bender et al. (2018b) limit themselves to only reporting these effects.

3.3.1 Primary results

Especially for the smooth terms, the interpretation of coefficients is hard and mostly even not very meaningful. Thus, Bender et al. (2018b) compute hazard ratios which are based on the predicted hazards from the model. While all covariates are neglected (set to their mean when predicting as suggested by Sylvestre and Abrahamowicz (2009)), only the nutritional associations are varied. Bender et al. (2018b) compute two separate hazards where different diet schemes have been assumed. They compute the ratio of these which is defined as:

$$e_j = \frac{\hat{\lambda}(j|Z_2(t))}{\hat{\lambda}(j|Z_1(t))} \tag{3.19}$$

where Z_1 and Z_2 represent the selected diets schemes for the comparison in e_j. The diet schemes can be any combination of the three daily diets C_I (lower), C_{II} (middle), and C_{III} (upper).

Using the model definition e_j can also be described by $exp(g_z) = exp(g_{z_1} - g_{z_2})$. e_j is then, though, a complex non-linear transformation demanding the use of the delta method. Thus, Bender et al. (2018b) make use of a trick to directly assess the difference of $Z_1 - Z_2$. One can simply define the exposure Z to be the exposure difference $Z_1 - Z_2$. $exp(Z)$ is then only a simple non-linear transformation where confidence intervals can easily be derived. Bender et al. (2018b) use the identity $X_Z = X_{Z1} - X_{Z2}$ for formula XX. That means that they use the design matrix of the difference of two distinct exposures as input already. As a result, this returns the difference between the cumulative effects straight from the model – with appropriate standard errors.

Bender et al. (2018b) are only interested in the impact of nutrition exposure from day 1 to 11 (this is, in fact, the period on which data is available). However, if one wanted to compare all (theoretically) possible combinations of daily diet schemes there would be $11^3 = 1331$ distinct diet schemes to be analysed. Bender et al. (2018b) are selective in their analysis and generally compare a diet scheme with a diet scheme which features slightly more (overall) nutritional intake. This way, Bender et al. (2018b) can limit themselves to **local** effect interpretations.

They formulate six relevant clinical comparisons, namely:

Table 3.1: Different comparisons of diets.

Comparison	Z_1	Z_2
A	Days 1-11: C1	Days 1-4: C1, Days 5-11: C2
B	Days 1-11: C1	Days 1-11: C2
C	Days 1-4: C1, Days 5-11: C2	Days 1-11: C2
D	Days 1-11: C1	Days 1-11: C3
E	Days 1-11: C2	Days 1-4: C2, Days 5-11: C3
F	Days 1-11: C2	Days 1-11: C3

These comparisons are graphically displayed below in figure 3.2.

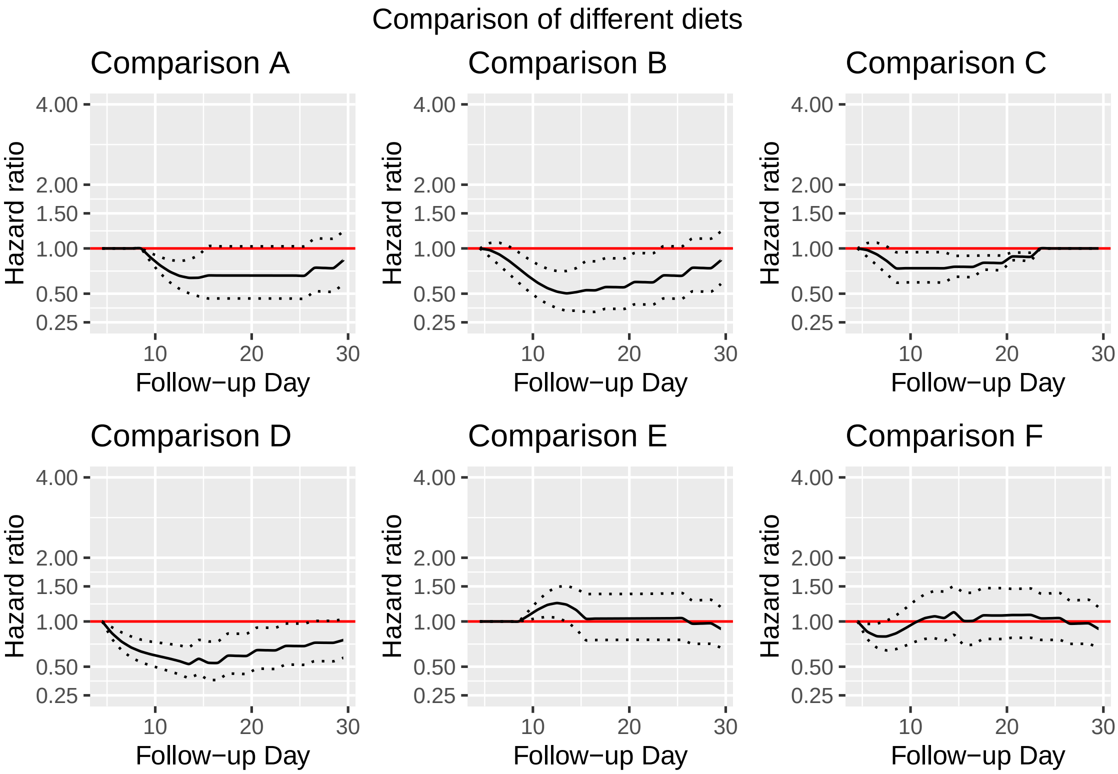

Figure 3.2: Comparison of different diets as in Bender et al. (2018).

Note that the graphic has been reconstructed. Within our approach we use – as previously outlined – slightly different data. However, the graph looks identical to Bender et al. (2018b). To be precise, the results of Bender et al. (2018b) are reproducible without using their **exact** code.

Hazard ratios below 1 mean that the ratio means that moving from the scheme Z_1 to Z_2 **decreases** the risk of dying. That means loosely speaking that hazard ratios smaller than 1 are associated with longer survival when higher-caloric nutrition is administered.

Panel 1 (upper left) depicts comparison A. The predicted hazard ratio decreases by moving to the higher-caloric diet and remains constant and a lower level for the complete follow-up. Z_1 and Z_2 being identical in the first four days causes the ratio to be fixed at 1 for these days. The upper confidence bands are smaller than zero for most points in time. This supports the interpretation that moving to higher caloric nutrition reduces the hazard.

Panel 2 (upper middle) depicts comparison B. The predicted hazard ratio behaves like in panel 1 without the identity of the first four days. The effect seems somewhat stronger, though. The hazard seems to decrease when moving to higher caloric nutrition.

Panel 3 (upper right) depicts comparison C. The predicted hazard ratio decreases significantly in the beginning. At some point – when the lagged association of the has completely disappeared (after 4 days being on a new diet) – the hazard ratio is constant at 1. Also in this scenario, the hazard seems to smaller if higher caloric nutrition was administered.

Panel 4 (lower left) depicts comparison D. The predicted hazard ratio is significantly smaller than 1 for the complete 30-day follow up. This comparison also indicates that higher caloric nutrition is associated with a decreased hazard.

Panel 5 (lower middle) depicts comparison E. The predicted hazard ratio is constant at 1 for the first days where both nutritional categories are identical between both schemes, Z_1 and Z_2. Then, the hazard ratio **increases** after the lagged effect of nutrition comes into effect. That means that here actually moving to the higher caloric scheme is associated with an increased hazard of dying. However, the increases are small in magnitude and not significant. This comparison does not indicate which of the two (lower vs. higher) schemes is to be preferred.

Panel 6 (lower right) depicts comparison F. The predicted hazard ratio is approximately constant at 1 for the complete 30-day follow up. Also, this panel does not give any evidence on which diet is preferable.

In sum, the first four comparisons, indicate that (marginally) higher-caloric nutrition should be preferred. However, this effect does not seem to apply if patients already receive a good amount (30-70 % of prescribed calories) of nutrition (as observable in the last two panels).

3.3.2 Clinical results

The clinical inference of the analysis of Bender et al. (2018b) – being an observational study – limited by construction. This is why Bender et al. (2018b) mainly focus on the interpretation of the marginal effects of nutrition. This is more likely to reflect the clinical reality where there are in principle many different constraints on the actual caloric intake. Hence, Bender et al. (2018b) are not to be understood as a recommendation to always give high-caloric nutrition. However, in the marginal case where a physician can adjust the caloric intake to some degree – following their analysis – it seems beneficial to exhaust the constraints as much as possible. Bender et al. (2018b) come to the overall conclusion though that (marginally (in fact not that marginal)) increasing the nutritional intake of critically ill patients on the ICU which require artificial nutrition leads c.p. to an increased survival time of these patients.

3.3.3 Methodological results

In a simulation study Bender et al. (2018b) investigate the model properties of the proposed model. From the modeled associations they simulate new data. Then, they re-model the association with deliberate misspecifications. They show that the PAMM which they used to model ELRA is robust to misspecifications (e.g. the lag-lead window). Additionally, they show that the proposed methods to quantify uncertainty usually operate very well.

However, the fixed lag-lead window size seems to possibly induce bias, and confidence intervals can have sub-nominal coverage.

3.4 Discussion and open research questions

As already outlined, the classical problems associated with observational studies also apply to Bender et al. (2018b). Confounding is a substantial problem in this context. Confounding is associated with an omitted variable problem which causes estimates to be biased. While Bender et al. (2018b) control for the diagnosis, the Apache II score, age, BMI, sex, and many more covariates, there will always be confounders that can either not be measured or have not been measured even if possible. One example is that in some cases the constraints on feeding will be very tight. These constraints can, e.g, be “procedures due to life-threatening complications” (Bender et al. (2018b)) which could not be anticipated by the Apache II score or the admission diagnosis. This means sometimes it is simply not possible to administer the intended caloric intake. In this case, this will eventually lead to biased population-based (or cluster-based) estimates of the nutritional intake. While this issue has to be considered when interpreting the results, Bender et al. (2018b) argue that they loosened it by investigating lagged (at least 4 days) effect for nutrition; this way recent medical conditions (which may lead to a very accelerated death) to not confound with the caloric intake.

Bender et al. (2018b) only investigate a single risk while – of course – it is most interesting to study what determines death in hospital. Bender et al. (2018b) treat hospital discharge as a non-informative event by censoring it administratively. However, this assumption is very strong and from a medical point of view, it makes sense to study if nutritional intake can also affect recovery times from a condition (i.e. speed up discharge time). Furthermore, competing risk survival analysis knows different models to combat the underlying association, from which Bender et al. (2018b) only investigated a single one, the Fine and Gray model for administratively censored data.

Chapter 4

Competing risks

In many cases in time-to-event analysis there is not only a single relevant possible event that could happen but multiple ones. The risks for these events mutually or separately affect the observed survival time. However, while an individual is under multiple risks, only one of all possible events can be observed when a competing risk setting is assumed.

Competing risks survival analysis deals with the fact that survival times for all risks but the observed one are missing. Consider an extreme example of two possible events: A and B. Assume that A is a very likely (some years of expected survival time) event while B is very rare (some million years of expected survival time). (E.g. consider the risk of a human dying and the risk of the whole solar system collapsing.) One would never observe one event (here B) because it is extremely unlikely compared to the other one (A). However, in the **absence** of the by far likelier risk, one would observe the unlikely event at some point in time.

In the case of Bender et al. (2018b) only the event 'death in hospital' is modeled. However, an individual is exposed to two separate risk processes. Namely, 'death in hospital' and 'hospital discharge'. While 'death in hospital' is a very specific event, 'hospital discharge' is vaguer. Even after discharge patients may die due to their previous condition. However, by explicitly modeling the discharge event instead of treating it as administratively censored data, one could better account for this specific event. In essence, it is interesting to study how patients' diets affect both risk processes. This is especially true if we expect that the processes are not independent of one another or if covariates either similarly or adversely affect the associated risks.

Modeling competing risks in the context of Bender et al. (2018b) is important to better understand the impact of nutrition on the patients. For example, while some diet may extend survival it could at the same time mean an extended hospital stay. This would result in an ambiguous interpretation of the effect of this diet. However, if survival time increases and discharge time decreases under a certain diet, this would rather support the interpretation of the diet effect in the first place.

Having motivated the modeling of competing risks in the context of Bender et al. (2018b), the aim of the rest of this chapter is an introduction to the modeling of competing risk which will be accessible to researchers and domain experts.

Beyersmann et al. (2011) regard competing risks models as special cases of multi-state models. Multi-state models deal with the modeling of transitions of individuals from one state to another. While in survival analysis typically there is just one transition, in multi-state modeling there are possibly many transitions of an individual. Furthermore, recurrent events are possible. However, multi-state-models are beyond the scope of this thesis and hence not further outlined.

For this thesis, a more helpful perspective (for the moment) is to view competing risks as latent failure times (which however is associated with multi-state modeling).

4.1 Competing risks as latent failure times

It is a prominent way to think about competing risks as latent failure times. Typically, competing risks involve a missing data problem: One can only observe the earliest event which happened.

So let Ξ be a set of all possible random variables $\{T_1, T_2, ..., T_K\}$ for K distinct events. These random variables reflect latent failure times. They are latent as one will only observe the minimal realisation of Ξ, namely $\tilde{T} = \min(\Xi)$. The remaining event times are missing. In a single risk context, all other risks than the one with the smallest latent failure time would be treated as if they are censored. For example, we track the survival time of cancer patients in a study. One of the patients, however, dies in an accident (i.e. due to another risk). The patient would be censored from the study event and the true event time will remain unknown.

The missing data situation imposes a fundamental problem in competing risk analysis. We do not know how strongly events are actually associated with one another. It is impossible to estimate the correlation structure of the events or event times accordingly.

4.2 Approaches to model competing risks

While latent failure times are a good way to think about competing risks, the standard approaches to model competing risks do not rely on the assumption of latent failure times (Beyersmann et al. (2011), p.50). Furthermore, the model is subject to the critique of creating an ‘artificial problem’ (Aalen (1987)) and hence not being suitable for real-world data analysis. There is great consent that the latent failure time model is not suitable for practical use. For a more detailed discussion, refer to Beyersmann et al. (2011) (p. 51).

Competing risks can be very well modeled without assuming latent failure times. However, the idea will play a substantial role in the motivation of different models.

Two dominant concepts build on single risk survival analysis.

1) Modeling of cause-specific hazards (Kalbfleisch and Prentice (2011)).

2) Modeling of subdistribution hazards (Fine and Gray (1999)).

Both modeling approaches can easily employ standard survival analysis estimators and models. Cause-specific hazards modeling aims to analyse the underlying risks in the absence of the other risks while subdistribution hazards incorporate these directly. In principle, the difference between cause-specific and subdistribution hazards results from different risk sets used for modeling. The risk set decreases for the cause-specific hazard model when an individual fails due to another risk than the one explicitly modeled. That means the individual is right-censored. When using subdistribution hazards these individuals **may** remain in the risk set for some positive time. Both cases only differ by the fact that implicitly different latent failure times are assumed.

Both approaches follow an identical procedure. Assume there is a set κ of K distinct risks due to which an individual can fail. We model all K risks separately after one another. For every single risk we ignore the existence of other risks (for both approaches) – we censor them. How they are censored, is very different between the cause-specific and subdistribution hazards model.

In both scenarios we *directly* model the hazard for a risk λ_q.

The cause-specific hazard for risk q is defined as:

$$\lambda_q^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, k = q | T > t)}{\Delta t} \tag{4.1}$$

In essence, this means that the marginal probability of failing by t due to risk q is estimated conditional on having survived so far. For the subdistribution hazards, however, we model:

$$\lambda_q^{sd}(t) = \lim_{\Delta t \to 0} \frac{P\left(t < T \leq t + \Delta t, k = q | T > t \cup (T \leq t \cap k = q^c\right)}{\Delta t} \tag{4.2}$$

In essence, we model this way the marginal probability of failing by t due to risk q conditional on having survived so far *and* not having failed due to any other risk which is not q prior to t.

The models do the very same however subject to different risk sets. There is a direct impact on the (non-parametric) estimation of the baseline hazard. Additionally, the maximum likelihood estimation of the regression coefficients (in a Cox model or PAMM) is indirectly affected, too. The estimation procedure remains unaffected, though.

4.3 Cause-specific hazards

Modeling cause-specific hazards is fairly simple. One only has to **ignore** all other risks when modeling the hazards for one specific risk. This is achieved by coding all failures due to different risks (from the one being modeled) as censored observations (Beyersmann et al. (2011)). In detail, one essentially does not change the event time but assumes censoring instead of the – in reality – observed competing event. Standard estimators such as the Nelson-Aaalen or Kaplan-Meier estimator can be used in a *cause-specific* manner. Furthermore, this approach allows the downstream use of standard survival models like the Cox model. However, some resulting estimates need to be dealt with great care (in terms of interpretation). Especially survival functions are not very meaningful: Cumulative incidences may sum up to a value larger than one.

Hence, we need to consider estimates which explicitly take other risks into account. So does the cumulative incidence function (CIF). The CIF can be directly modelled via the subdistribution hazard model (which we will see later) or indirectly computed via the cause-specific hazards model. From the cause-specific estimates one constructs the CIF as follows (Haller (2014)):

$$\underline{F}_k(t) = \int_0^t \lambda_k(s) S_{ov.}(s) \mathrm{d}s = \int_0^t \lambda_k(s) \exp\left(-\sum_{l=1}^{K} \Lambda_l(s)\right) \mathrm{d}s \tag{4.3}$$

This means that the CIF weights the overall survival by the cumulative hazard of interest. All measures of interest can be estimated (for discrete points in time) easily with the cause-specific model.

This estimator function is a special case of the Aalen-Johansen estimator for transition probabilities in multi-state models (Haller (2014)). However, as the CIF is a non-linear transformation, model inference is not straight-forward. The variance of the CIF is unknown. One solution is to use the delta method. The approximated variance (from the delta method) for the CIF is (Iljon (2013)):

$$\begin{aligned}\operatorname{Var}\left(\underline{\hat{F}}_k(t)\right) = \sum_{t_j \leq t} \Bigg\{ & \left[\underline{\hat{F}}_k(t) - \underline{\hat{F}}_k\left(t_j\right)\right]^2 \frac{d_j}{n_j\left(n_j - d_j\right)} + \left[\hat{S}\left(t_{j-1}\right)\right]^2 \frac{n_j - d_{kj}}{n_j^3} \\ & -2\left[\underline{\hat{F}}_k(t) - \underline{\hat{F}}_k\left(t_j\right)\right]\left[\hat{S}\left(t_{j-1}\right)\right] \frac{d_{kj}}{n_j^2} \Bigg\}\end{aligned} \tag{4.4}$$

where d_j is the number of failures in the interval t_{j-1} until t_j and n_j is the magnitude of the risk set at point t_j.

Using the cause-specific hazard approach we obtain the following measures of interest with (approximate) standard errors:

1) Cause-specific hazard and cumulative hazard functions. They are interesting to study the risk processes themselves. Cause-specific hazards are still **marginally** interpretable.
2) CIFs. They are interesting to study as they allow a survival (probability) based interpretation.
3) The overall survival function can be estimated from the cause-specific hazards, too.

When employed in a regression context (e.g. Cox model) the cause-specific maximum likelihood estimates are to be interpreted in the absence of all other risks. This is the major aspect of the critique of the cause-specific hazards model. Since this model neglects the interdependence of all other risks when model coefficients are estimated, these coefficients may be *biased* w.r.t. the CIF. This originates in the fact that cumulative incidences may add up to a value larger than one.

However, the term of biasedness is not placed properly within this discussion. The final section of this chapter outlines that there is an ongoing discussion on whether or not to model cause-specific hazards. Referring to the cause-specific estimates as biased implies that the cause-specific model is the wrong model and the subdistribution hazards model should be used instead. This, though, cannot be generalised and must be evaluated concerning the underlying data situation.

4.4 Subdistribution hazards

Subdistribution hazards are a smooth approach to directly model CIFs (without approximated standard errors). This is because the CIF is essentially the $1 - S_k(t)$ estimate which is determined by the subdistribution hazards instead of the cause-specific hazards. Furthermore, if used in a regression context this approach can facilitate estimates which account for the interdependence of the different risks. Essentially, subdistribution hazards can be determined by the same estimators as the cause-specific hazards or hazards in a single-risk environment (as seen in Beyersmann et al. (2011)). Thus, for regression models likelihood-based inference is possible.

The idea of the subdistribution framework is to **better** assess the risk set over the whole survival time. When working with the cause-specific hazards model, we simply treated competing events as censored events. We do the same for the subdistribution hazards model. However, we do **not** use the event time of the competing risk as censorship time. We **guess** a suitable censorship time. The Fine and Gray model (Fine and Gray (1999)) is the most prominent model which does this. However, virtually any approach (especially imputation or prediction), which succeeds predicting the expected censorship time, is – in practice – possible.

The following sections outline the Fine and Gray model as the standard model for subdistribution hazards.

4.4.1 Fine and Gray model

Traditionally, the modeling of subdistribution hazards is closely tied to the model proposed by Fine and Gray (1999). They motivate the model via the CIF. The CIFs can be estimated consistently and directly via their model. This direct modeling implies two advantages:

1) Possible covariate effects reflect the association with the CIF and hence account for all possible risks.
2) Inference is typically straight forward as the CIF can be directly inferred from the model.

The model of Fine and Gray (1999) aims to assess the previously outlined subdistribution hazard rate:

$$\lambda_q^{sd}(t) = \lim_{\Delta t \to 0} \frac{P\left(t < T \leq t + \Delta t, k = q | T > t \cup (T \leq t \cap k = q^c\right)}{\Delta t} \tag{4.5}$$

The CIF for q can be inferred from this model by:

$$F_q(t) = 1 - \exp\left[- \int_0^t \lambda_q^{sd}(u) \mathrm{d}u \right] \tag{4.6}$$

From this representation of the CIF, we can see an important analogy. The subdistribution hazard directly affects the CIF. This works perfectly analogously to the single risk model where the hazard rate affects the survival function. This analogy – the direct link between these two – can be exploited especially for the estimation of covariate effects.

Effectively, the subdistribution hazards model is estimated by a modification of the risk set. The risk set for the subdistribution hazards approach considers all individuals which have not failed due to risk $k = q$ by time t. All these individuals are at risk at time t. Hence, the cause-specific and the subdisributional hazard for one individual are equivalent until one event is observed. From this moment on, $\lambda_q^{sd}(t)$ and $\lambda_q^{cs}(t)$ are no longer equivalent for any $q \in K$; $\lambda_q^{sd}(t)$ must be smaller than $\lambda_q^{cs}(t)$.

Fine and Gray (1999) use a weighted likelihood-based approach to estimate the parameters of the model. When embedding the framework of Fine and Gray (1999) to the Cox model, the partial likelihood is weighted. The weighting is necessary to adjust the risk set throughout the survival period. This is necessary because in practice it turns out to be problematic that the subdistribution hazard has probability mass at infinity (Beyersmann et al. (2011)).

In general – in line with Haller (2014) and Fine and Gray (1999), we outline three cases here:

1) Completely observed data

2) Administrative Censoring

3) Incomplete data

The terms complete and incomplete refer in this setting to truly censored data. That means if there are (right) censored individuals, the data is incomplete. All three cases adjust the risk set in such a manner that it better reflects reality. For incomplete data, censorship has to be taken into account explicitly when doing so.

4.4.1.1 Completely observed data

This case is essentially the simplest one. As there is no actual censoring existing, one can ignore censoring or the censor distribution in this case. For the modeling of $\lambda_q^{sd}(t)$ we need to adapt the risk set only slightly. The adaptation of the risk set at the time of the failure of the i-th individual looks like the following.

$$R_i = \{j : (t_j \geq \tilde{t}_i) \cup (t_j \leq \tilde{t}_i \cap k_j \neq q)\} \tag{4.7}$$

This means an individual j is part of the risk set at t_i if its survival time t_j is larger than t_i or t_j is smaller than t_i but the event k_j was not of the same kind as for i namely q

We only allow individuals in the risk set which are still at risk at t_{ki}. This includes all individuals who did not fail yet due to any risk and who failed from a risk different from k.

From a practical point of view, this risk set is achieved by coding events all different from k as censoring with a maximum censor time. This maximum could be the maximally observed event time. Technically, this is very similar to artificially created administrative censoring.

4.4.1.2 Administrative censoring

In this spirit, we present a conceptually slightly more complex case: administrative censoring. Here, the potential censoring time is always observed, namely the maximum follow-up. We denote the censor time for individual i as c_i.

The risk set at the time of the failure of the i-th individual is:

$$R_{\tilde{t}_i} = \{j : (min(t_j, c_j) \geq \tilde{t}_i) \cup (t_j \leq \tilde{t}_i \cap k_j \neq q \cap c_j \geq t_i)\} \tag{4.8}$$

This means an individual j is part of the risk set at t_i if its survival time t_j or administrative censor time c_j (whichever happens first) is is larger than t_i. An individual is also a part of the risk set if t_j is smaller than t_i but the event k_j was not of the same kind as for i namely q *and* the censoring time c_j is larger than t_i.

Just like for the complete data case the risk set consists of all individuals at risk or not at risk but failed due to another risk than modeled. Both cases can be dealt with identically from a practical point of view. All individuals who failed due to another risk than q are censored at the time of administrative censoring.

Note that this procedure is only feasible if administrative censoring is the only censoring that exists.

4.4.1.3 Incomplete data

Incomplete data impose a significant problem on the procedures just presented. If the only way to leave the study is via one of the competing events (or the end of the study), the only possible censorship time is known. For incomplete data these possible censorship times are unknown. Fine and Gray (1999) make use of the information in the data on the censorship distribution $G(t)$ and model it explicitly. They use this distribution to weight all individuals in the risk set. Weights are determined by inverse probability of censoring weighting (IPCW) as proposed by Rotnitzky and Robins (2005).

Haller (2014) defines the time-dependent weights generated by IPCW in the following manner, where w_i represents the weight of individual i:

$$w_i(t) = r_i(t) \frac{\hat{G}(t)}{\hat{G}(min(t_i, t))} \tag{4.9}$$

$r_i(t)$ is an indicator for individual i to be still under risk at t.

$$r_i(t) = I(c_i \geq min(t_i, t)) \tag{4.10}$$

It is 1 if i is still under risk at t or the individual failed due to another risk than q. It is 0 if she has been censored before t. $\hat{G}(t)$ and $\hat{G}(min(t_i, t))$ are typically (in the standard models) the Kaplan-Meier estimators.

One can easily observe the following properties:

1) Censored individuals disappear from the risk set with a weight of 0.
2) Individuals under risk remain in the risk set with a weight of 1.
3) Individuals not under risk but also not censored remain in the risk set with a weight smaller than 1.

To be precise, the weight w_i equals to $\frac{\hat{G}(t)}{\hat{G}(t_i)}$ in the last case. $\hat{G}(t)$ is the Kaplan-Meier estimate evaluated at t and $\hat{G}(t_i)$ the Kaplan-Meier estimate at the observed event time t_i. $\hat{G}(t_i)$ is static and $\hat{G}(t)$ time-dependent. The weight is diminishing over time with decreasing value of $\hat{G}(t)$.

The weights are used for the estimation of the baseline hazards and later on in the score function for the ML coefficient estimation.

This approach demands a two-staged procedure. First of all, the Kaplan-Meier survival function for the censorship distribution needs to be estimated. From this point on, a weighted regression can be performed to derive non-parametric and parametric quantities.

Nevertheless, the use of the Kaplan-Meier estimator for all individuals suggests that censoring is independent of the covariates. In many cases, this is not plausible. Then, the censor distribution $G(t)$ can be predicted by a survival model itself for each individual.

4.5 Modeling covariate effects

For both models, the data needs to be transformed in the previously described manner. Then standard models (e.g. Cox PH) can be used for modeling.

Theoretically, the modeling of any kind of effect feasible for single risk analysis can be employed in **both** models. However, the subdistribution hazards model in combination with time-dependent covariates imposes a potential problem. We use a censorship time which is larger than the observed one. For the whole additional time resulting from this, the time-dependent covariate set is unknown. This naturally imposes another missing data problem which can be solved by assumptions or imputation (/prediction). For example, Beyersmann et al. (2011) proposes the forward carrying of the last observed covariate. For an overview of time-dependent covariates in the Fine and Gray model consider Austin et al. (2020). Their literature review can be summarised by the following: Many applied researchers are not aware of the missing data problem in the Fine and Gray model with time-dependent covariates. Solving this problem is still the subject of ongoing research.

ELRA, as presented in the previous section, are an interesting effect: it is a smooth multivariate time-varying effect. Incorporating these sorts of effects implies that a more simple effect can also be facilitated. Bender et al. (2018b) show that ELRA can be easily used in PAMMs for the Fine and Gray model for administratively censored data (which is a special case of the right-censored case). The application of these for cause-specific hazards is trivial. While PAMMs are especially good at incorporating these sorts of events, there also have been recent attempts to use the Cox PH model for complex effects of this style. For example, Danieli and Abrahamowicz (2019) implemented cumulative effects into an extended Cox PH framework for cause-specific hazards. They build on Lunn and McNeil (1995) and Belot et al. (2010) to preprocess the data similarly as performed for PAMMs to facilitate the longitudinal structure. The ELRA specification chosen by them is closely related to Gasparrini (2014).

4.6 Summary

Competing risks analysis is all about assumptions on the risk set. Either the risk set is immediately updated when a competing risk occurs (cause-specific hazards model); or the risk set is artificially extended if competing risks occur (Fine and Gray model for complete data and administrative censoring); or the risk set is newly constructed based on the information on the censorship distribution which is available (Fine and Gray model for incomplete data).

4.7 Cause-specific vs subdistribution hazards – a review

While both models find broad application in studies in survival analysis, there is a richness of opinions which model is the **better** one. As we outlined in this thesis, the main degree of freedom in modeling hazards (cause-specific vs. subdistribution) is the assumed censorship. Modeling cause-specific hazards explicitly assumes that censoring is non-informative. Testing this assumption is not possible (Tsiatis (1975)). However, in the competing risk context, it is reasonable to assume that it may be violated. In this case, estimates can be biased. Frequently, this is the motivation for researchers to move to the Fine and Gray model. However, when moving away from cause-specific hazards something very different is modeled. Cause-specific hazards and subdistribution hazards

cannot be interpreted in the same way. Thus, a direct comparison of coefficients between these models is most of the time not meaningful.

Austin and Fine (2017) summarise the conflict by the following: “The use of a cause-specific hazard model allows one to estimate the effect of the intervention on the instantaneous rate of occurrence of the event of interest in subjects who are currently event-free. The use of a subdistribution hazard model allows one to estimate the relative effect of the intervention on the cumulative incidence function.” This means that modeling cause-specific hazards is a better approach for the investigation of causal effects. This is in line with the observation of Beyersmann et al. (2011) that cause-specific hazards only define the risk process. Lau et al. (2009) also suggest using cause-specific hazards for causation while they favour subdistribution hazards for predictive tasks. According to the authors, subdistribution hazards are particularly useful when assessing the **probability** of failing due to a specific risk at a given point. This is because the direct modeling of the CIF incorporates **all** separate risk processes. In fact, Fine and Gray (1999) did not claim that modeling subdistribution hazards would result in estimating causal effects. Their focus was a way to estimate CIFs; thus, this section does not oppose Fine and Gray (1999), but even rather supports their initial intent.

Both models have their advantages and always the context should be considered where these are applied: cause-specific hazards fit etiologic research and subdistribution hazards predicting the observed rate of occurrence. In total, we agree with Beyersmann et al. (2011) and Grambauer et al. (2010a) who suggest using both model vis-à-vis and interpret them jointly.

Chapter 5

Competing risks for PAMMs

This chapter incorporates the scientific novelty of this thesis. The previous chapter examined competing risks with the aim to abstract the concepts in such fashion that they can be easily applied to PAMMs. This is consequently done in this chapter for both competing risks models presented in this thesis, the cause-specific hazard model, and the subdistribution hazard model. The principal aim of this chapter is to transfer both models to the PAMM context. Thus, solving problems that are not related to PAMMs (e.g. time-dependent covariates in the subdistribution hazard model) are not the primary focus of this chapter. This chapter is supported by `R` code which illustrates the implementation of competing risks PAMMs in the `pammtools` package.

All concepts derived from the previous chapter have been presented on such an abstract level that it is not necessary to change anything about the general modeling procedure of PAMMs. As usually when working with PAMMs the preprocessing is already an essential modeling part. Modeling competing risks PAMMs is modular and does not require (much) new theory.

In this section, the implemented methods are illustrated using the `sir.adm` data set from the `mvna` package. The data features data on “pneumonia status on admission for [...] ICU [...] patients”, and is “a random sample from the SIR-3 study”. Next to pneumonia status, the data covers information on the survival status, the survival time, the patients’ age, and their sex.

```r
library(mvna)
data("sir.adm")
sir_adm <- sir.adm
head(sir_adm)
```

id	pneu	status	time	age	sex
41	0	1	4	75.34	F
395	0	1	24	19.17	M
710	1	1	37	61.57	M
3138	0	1	8	57.88	F
3154	0	1	3	39.01	M
3178	0	1	24	70.28	M

The event type is stored in `status`.

```r
table(sir_adm$status)
```

0	1	2
14	657	76

We observe that it is an `integer` which is either 0, 1, or 2. Considering the documentation of the data set we learn more about the type of events and the other covariates. Refer to the Appendix to access the documentation or use `?sir.adm`.

We see that we have a competing risks situation (discharge vs. death) with right-censored data. In the following, we outline different competing risks PAMMs based on the insights from the previous chapter. We focus on the cause-specific hazards and subdistribution hazards model. We illustrate the model fitting with the `sir.adm` data.

5.1 Cause-specific hazards PAMMs

As outlined in the previous section, competing events simply need to be re-coded as censoring events in the cause-specific framework. The event time remains the same. When performed for all distinct risks, this eventually results in as many data sets as there are risks.

For each risk, the modeling as a PAMM (i.e. the pipeline of preprocessing and model fitting) is performed separately in the same manner. A cause-specific PAMM for one competing is equivalent to the single-risk PAMM from chapter 2.

Also, the GAMM modeling (via `gam()`) is the very same as in the single risk case for each distinct risk. Having studied the previous chapters of this thesis this procedure is trivial and hence not further outlined here.

One only needs to loop over all different risks and perform the same fitting procedure – consisting of transformation to a piece-wise exponential data frame and the actual modeling of a `gam` – for each of these risks. While this could be done with low coding effort by the user herself, we provide the convenience functions `as_ped_cr_cs()` and `pem_cr_cs()` and `pam_cr_cs()`. `as_ped_cr_cs()` wraps `pammtools::as_ped()` and recodes the event type beforehand. It returns a list of `ped` data sets. `pem_cr()` wraps `stats::glm()` for all different risks and their respective data sets. It returns a list of models. `pam_cr()` wraps `mgcv::gam()` for all different risks and their respective data sets. It returns a list of models.

One obtains a cause-specific competing risks hazards PAMM with the following code:

```
cut_points <- c(0:40, seq(45, 75, by = 5), seq(80, 100, by = 10), 120)
ped_cs <- as_ped_cr_cs(sir_adm, Surv(time, status) ~ ., cut = cut_points,
                       id = "id")
pam_cs <- pam_cr(ped_status ~ s(tend) + pneu + s(age) + age:tend +
                   s(sex, bs = "re") - 1, ped = ped_cs, family = poisson(),
                 offset = offset)
```

The cut points are manually crafted here. We are only interested in the first 120 days as after this day there are very few incidents so that results will be anyways very questionable. The `ped_cr` object is a `list` of `ped` data sets for each distinct risk. Here it has length 2. The `pam_cr()` function is used in the same way as in the single risk scenario `gam()`.

This competing risk PAMM has a smooth baseline, a smooth, linearly time-varying effect of age, and a random effect of the sex. Pneumonia is linearly modeled. There is no intercept.

The resulting models have `summary` and `print` methods. However, `R` summaries turn out to be lengthy. We only report the resulting coefficients here.

```
summary_pam_cs <- summary(pam_cs)
data.frame(discharged = summary_pam_cs[[1]]$p.coeff,
           dead = summary_pam_cs[[2]]$p.coeff)
```

	discharged	dead
pneu	-1.1026	-0.0578
age:tend	-0.0004	0.0004

For interpretation, we aim to focus on the covariate `pneu` – the indicator for pneumonia. We discover, that c.p. having pneumonia significantly negatively affects the hazard of being discharged. However, for the competing status `dead` we do not observe any effect.

Not all methods for `gams` have been wrapped. So, for example, to use the `plot()` method for `gams` one has to call it on every single element of the list. One can also call `summary()` on every single element of the list.

We also provide convenience functions to compute measures inferred from the model:

- `add_hazard_cr()` wraps `add_hazard()` and computes the resulting hazard estimates (via the predict method for the `gam`).
- `add_cumu_hazard_cr()` wraps `add_cumu_hazard()` and computes the resulting cumulative hazards estimates.
- `add_surv_prob_cr()` wraps `add_surv_prob()` and computes **cause specific** survival estimates. Note that these estimates are not necessarily meaningfully interpretable (as outlined in the previous chapter).
- `add_cif()` computes the cumulative incidence function which eventually wraps `add_hazard()` and `add_cumu_hazard()` and combines the input in the fashion explained in the previous chapter.

All functions take the model object and return – in the same style as in standard `pammtools` – the associated estimates for all different risks (if applicable).

We show how some of the functions work in the following code example.

First, we prepare a PAMM which only models the baseline.

```
pam_cs_base <- pam_cr(ped_status ~ s(tend), ped = ped_cs, family = poisson(),
                      offset = offset)
```

Then, we create a new data frame on which we want to predict.

```
interval_df <- int_info(ped_cs[[1]])
head(interval_df)
```

tstart	tend	intlen	intmid	interval
0	1	1	0.5	(0,1]
1	2	1	1.5	(1,2]
2	3	1	2.5	(2,3]
3	4	1	3.5	(3,4]
4	5	1	4.5	(4,5]
5	6	1	5.5	(5,6]

Subsequently, we predict cumulative hazards and append them to the data frame.

```r
interval_df <- interval_df %>%
  add_cumu_hazard_cr(pam_cs_base, ci = TRUE, ci_type = "sim")
head(interval_df[, 5:8])
```

interval	1_cumu_lower	1_cumu_upper	1_cumu_hazard
(0,1]	0.0410	0.0579	0.0496
(1,2]	0.0886	0.1212	0.1053
(2,3]	0.1438	0.1908	0.1675
(3,4]	0.2067	0.2667	0.2365
(4,5]	0.2780	0.3470	0.3121
(5,6]	0.3519	0.4339	0.3936

```r
head(interval_df[, c(5, 9:11)])
```

interval	2_cumu_lower	2_cumu_upper	2_cumu_hazard
(0,1]	0.0014	0.0061	0.0031
(1,2]	0.0033	0.0123	0.0066
(2,3]	0.0056	0.0185	0.0107
(3,4]	0.0087	0.0248	0.0154
(4,5]	0.0127	0.0316	0.0208
(5,6]	0.0171	0.0391	0.0269

```r
ggplot(interval_df, aes(x = tend, y = `1_cumu_hazard`)) +
  geom_line() +
  geom_line(aes(x = tend, y = `1_cumu_lower`), linetype = "dotted") +
  geom_line(aes(x = tend, y = `1_cumu_upper`), linetype = "dotted") +
  geom_line(aes(x = tend, y = `2_cumu_hazard`), col = 2) +
  geom_line(aes(x = tend, y = `2_cumu_lower`), linetype = "dotted", col = 2) +
  geom_line(aes(x = tend, y = `2_cumu_upper`), linetype = "dotted", col = 2) +
  ggtitle("Cumulative baseline hazards (120 days)",
          subtitle = paste("Discharge (black) vs. Death (red) with 95%",
                           "confidence bands (dotted lines).")) +
  xlab("Days") + ylab("Cumulative Hazard")
```

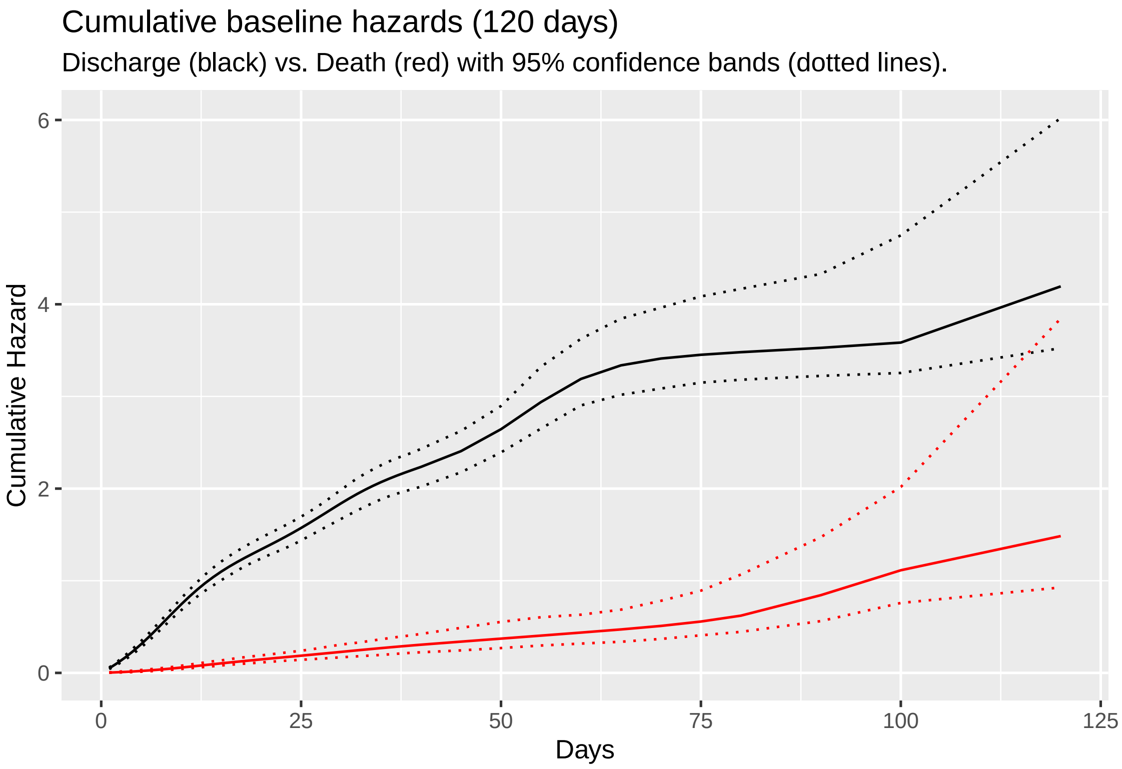

Figure 5.1: Cause-specific cumulative hazards estimates using a PAM.

The cumulative incidence function can be created in the same manner. For this function note that we currently only provide simulated confidence intervals for cause-specific hazards. Modeling the CIF for cause-specific hazards additionally (compared to other conv. functions) requires the providing of the `ped` object used for the model fitting.

```
interval_df <- int_info(ped_cs[[1]])
interval_df <- interval_df %>% add_cif(pam_cs_base, ped_cs, ci = TRUE)
head(interval_df[, 5:8])
```

interval	1_cif	1_cif_lower	1_cif_upper
(0,1]	0.0482	0.0393	0.0586
(1,2]	0.0994	0.0834	0.1177
(2,3]	0.1532	0.1325	0.1767
(3,4]	0.2092	0.1845	0.2366
(4,5]	0.2659	0.2384	0.2962
(5,6]	0.3215	0.2934	0.3553

```
head(interval_df[, c(5, 9:11)])
```

interval	2_cif	2_cif_lower	2_cif_upper
(0,1]	0.0030	0.0016	0.0057
(1,2]	0.0064	0.0036	0.0113
(2,3]	0.0100	0.0058	0.0170
(3,4]	0.0140	0.0082	0.0227
(4,5]	0.0181	0.0109	0.0279
(5,6]	0.0222	0.0138	0.0331

```r
ggplot(interval_df, aes(x = tend, y = `1_cif`)) +
  geom_line() +
  geom_line(aes(x = tend, y = `1_cif_lower`), linetype = "dotted") +
  geom_line(aes(x = tend, y = `1_cif_upper`), linetype = "dotted") +
  geom_line(aes(x = tend, y = `2_cif`), col = 2) +
  geom_line(aes(x = tend, y = `2_cif_lower`), linetype = "dotted", col = 2) +
  geom_line(aes(x = tend, y = `2_cif_upper`), linetype = "dotted", col = 2) +
  ggtitle("Cumulative incidences (120 days)",
          subtitle = paste("Discharge (black) vs. Death (red) with 95%",
                           "confidence bands (dotted lines).")) +
  xlab("Days") + ylab("Cumulative incidences")
```

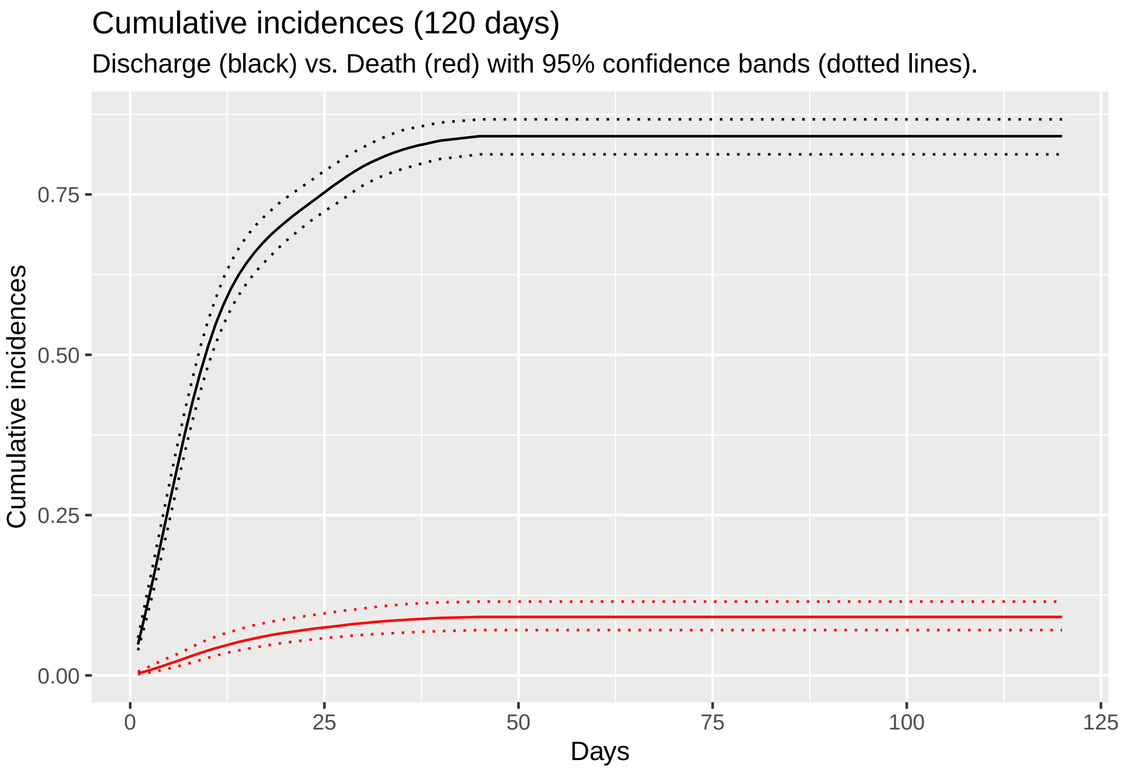

Figure 5.2: CIF estimates using a cause-specific hazards PAM.

5.2 Subdistribution hazards PAMMs

The modeling approach for the subdistribution hazards model is less straight forward. For `pammtools` we embedded the Fine and Gray model with its three sub-models. While the models for complete data and administrative censoring are very straight forward, the model for incomplete data requires more effort.

The intuition of the Fine and Gray model is to find smarter or better guesses for the censorship times for competing events. Recall that an appropriate censoring time for data with no right-censoring is the maximum observed event time. Data with administrative censoring are appropriately censored at the end of the study. For incomplete data, a weighted risk set has been suggested by Fine and Gray (1999).

As outlined in chapter 3, time-dependent covariates are not easily implemented into the subdistribution hazards frameworks as proposed by Fine and Gray (1999). In the absence of a better solution, our approach is the “carry-on” method (this is the default in `pammtools`): The last observed value of a time-dependent covariate is carried on until the end of the hypothetical follow-up. However, we are not at all convinced by this method and propose further research on this topic.

5.2.1 No censoring and administrative censoring

Complete data (no censoring) and administrative censoring is almost identically treated in practice. In both cases, one carries on the risk set as long as possible for observations who failed due to a different risk. While for complete data an appropriate censoring time would be infinity, in practice the *maximally observed event time* has identical implications. For the administrative censoring, the *maximally observable event time* is the end of the study. These two cases are easily handled. Before the transformation to a piece-wise exponential data frame there are two necessary steps (where for cause-specific hazards there was just one):

- Recode the competing event as censoring.
- Recode the event time to either t_{max} (no censoring) or the a priori defined administrative censor time.

Like for cause-specific hazards, we finally return a list of piece-wise exponential data sets. However, these data sets are **not** anymore equivalent to the true data sets but incorporate the assumptions by Fine and Gray (1999). From here on, the same methods (e.g. `pam_cr()` or `gam()`) as for cause-specific hazards can be applied.

5.2.2 Incomplete data

For incomplete data, Fine and Gray (1999) suggest the use of a weighted score function. However, weights – as used in a cox model – do **not** directly translate to weights in the equivalent Poisson representation. This means that one cannot directly transfer the Fine and Gray model to PAMMs. Still, we can make use of their idea to “guess” survival times to increase the risk set stochastically. We suggest to indirectly make use of survival estimates in our model to facilitate the idea of Fine and Gray (1999).

The proposition of this thesis is the following: First, the competing event is recoded to a censoring event for all distinct events. Then, the event time of this censorship is predicted based on the information available. This means that we model the survival function for censoring (in a cause-specific fashion) and use it to predict (or simulate) survival times. We make use of the piece-wise exponential assumption and predict the survival time (for being censored) for each individual conditioned on the fact that the individual already survived up to now. We use the predicted hazard for being censored then when there was a competing event $\hat{\lambda}_i^{cens}(t_i^k)$ (where k represents the event that occurred for i) and simulate a (piece-wise exponential) survival time for it. We do so by plugging the hazard into an exponential distribution and drawing from it a single time.

$$\tau_i \sim \mathcal{E}xp(\hat{\lambda}_i^{cens}(t_i^k)) \tag{5.1}$$

$$\tilde{t}_i^{cens} = t_i^k + t_i^{inc} \tag{5.2}$$

where t_i^{inc} is a realisation of τ_i.

Note that for a PAMM, we could also use the end of the interval which t_i^k lies in instead of t_i^k.

This indirect approach solves one core problem of Fine and Gray (1999), namely that the classical Fine and Gray model only works well in the presence of independent censoring. It is only possible to work with stratified estimates when the Fine and Gray model is applied. We can now **easily** loosen the assumption of independent censoring as we allow covariates to affect censoring as well. This is very well facilitated in our framework because we semi-parametrically estimate censorship hazards. However, this comes at the cost of the parametric assumption of a piece-wise exponential censorship distribution.

We also embedded the Fine and Gray model (or rather an analogon) into `pammtools`. As for incomplete data, we need to make use of a PAMM in the preprocessing already, the preprocessing procedure will be much more computationally expensive than for the other models.

We offer two functions from the `as_ped()` family:

- `as_ped_cr_sh()` for no and administrative censoring. This function wraps again `as_ped()`. Nevertheless, before that, the censoring time is re-coded based on the argument `max_time` which defaults to the maximum of the observed event times if not provided. Missing time-dependent-covariates for competing risks are carried on for all risk sets.
- `as_ped_cr_cens()` for incomplete data. The function implements exactly the procedure described in the section for incomplete data. Missing time-dependent-covariates for competing risks are carried on for all risk sets. Both cases, independent censoring (default) and dependent censoring (via providing a censoring formula) are covered by the function.

Like the previous models, the final object of these functions returns a list of `ped` data frames. Each element is the modified (so that the assumptions by Fine and Gray (1999) are met) data set for every single risk.

5.2.3 Modeling subdistribution hazards

The actual modeling is more straight forward. As already at preprocessing time we incorporate everything that distinguishes the different models, we can still simply use GAMMs with Poisson likelihood for modeling. To account for the data sets coming in a list of data sets, one can conveniently use the previously presented functions `pem_cr_cs()` or `pam_cr_cs()`.

However, if the user is only interested in the analysis of a single risk, we recommend to simply fit a `gam` on the data frame of interest out of the list returned by the function of the `as_ped_cr` family.

Concerning our example `sim.df` data it is important to note that there **is** censoring but there are only 14 censored individuals so that we decide that we can neglect this at this point. One may have noticed that we already performed some sort of administrative censoring before limiting the data to the first 120 days. Now, we are more restrictive and do as if the study ended after 90 days. Then, we compute the subdistribution hazards model for administratively censored data. `max_time = 90` instructs the function to recode all competing event times with `90`. We apply the same model as before. Now, however, the hazards which are modeled are subdistribution ones.

```
cut_points <- c(0:40, seq(45, 75, by = 5), 80, 90)
ped_sh <- as_ped_cr_sh(sir_adm, Surv(time, status) ~ ., cut = cut_points,
                       max_time = 90, id = "id")
pam_sh <- pam_cr(ped_status ~ s(tend) + pneu + s(age) + age:tend +
                   s(sex, bs = "re") - 1, ped = ped_sh, family = poisson(),
                 offset = offset)
summary_pam_sh <- summary(pam_sh)
data.frame(discharged = summary_pam_cs[[1]]$p.coeff,
           dead = summary_pam_cs[[2]]$p.coeff)
```

	discharged	dead
pneu	-1.1026	-0.0578
age:tend	-0.0004	0.0004

Like in the cause-specific model, there is a significantly negative impact of `pneum` on the subdistribution hazard for discharge. However, unlike the previous model, we also measure a significantly positive effect of having pneumonia on the subdistribution hazard of death.

To illustrate the model for incomplete data we select a data set that has a higher proportion of censored individuals. The `fourD` data set from the `etm` package seems to fit well. The documentation (accessible via `?fourD`) can be found in the Appendix. The data description features the following information: "Data from the placebo group of the 4D study. This study aimed at comparing atorvastatin to placebo for patients with type 2 diabetes and receiving hemodialysis in terms of cardiovascular events. The primary endpoint was a composite of death from cardiac causes, stroke and non-fatal myocardial infarction. Competing event was death from other causes." The data holds information on the patient's sex, age, medication, survival time, and survival status, such as their treatment status.

```
library(etm)
data("fourD")
table(fourD$status)
```

0	1	2
264	243	129

For this data, we only focus on the modeling of the (smooth) baseline.

```
fourD$time[fourD$time > 6L] <- 6L ## one observation with time 6.001
cut_points <- seq(0, 6, by = 0.1)
ped_cens <- as_ped_cr_cens(fourD, Surv(time, status) ~ .,
                           cut = cut_points, id = "id")
```

The provided code simulates the censorship time for competing events independently from covariates. In order to facilitate this, one could try the following code:

```
censor_formula <- as.formula("ped_status ~ s(tend) + age + sex + treated")
ped_cens <- as_ped_cr_cens(fourD, Surv(time, status) ~ ., cut = cut_points,
                           id = "id", censor_formula = censor_formula)
```

For the subdistribution hazards PAMM (all three sub-models), we can use (almost) the same post-processing functions as for the cause-specific model. To predict from the model we can use:

- `add_hazard_cr()`: Technically, this function is identical to the one used for the cause-specific model. However, the interpretation of the hazards is now different as they are subdistribution hazards.
- `add_cumu_hazard_cr()`: The same is true for this function.
- `add_surv_prob_cr()` is also identical to the implementation for cause-specific hazards. However, now it makes more sense as every single risk's "survival" function represents its contribution to the CIF.
- `add_cif()` computes the cumulative incidence function via `add_surv_prob()`. `add_cif()` has two methods, one for subdistribution and one for cause-specific hazards PAMMs. The user simply needs to call the function, though.

We only show the `add_cif()` function (`add_cif.sh()` to be precise) here. One can also supply a `ped` object like for the `add_cif.cs()` method to this method but it will eventually be ignored.

```
pam_cens_base <- pam_cr(ped_status ~ s(tend), ped = ped_cens,
                        family = poisson(), offset = offset)
interval_df <- int_info(ped_cens[[1]])
interval_df <- interval_df %>% add_cif(pam_cens_base, ci = TRUE)
head(interval_df[, 5:11])
```

interval	1_cif	1_cif_upper	1_cif_lower	2_cif	2_cif_upper	2_cif_lower
(0,0.1]	0.0103	0.0065	0.0164	0.0053	0.0034	0.0081
(0.1,0.2]	0.0210	0.0139	0.0319	0.0106	0.0070	0.0160
(0.2,0.3]	0.0322	0.0221	0.0470	0.0160	0.0108	0.0236
(0.3,0.4]	0.0437	0.0311	0.0618	0.0214	0.0147	0.0311
(0.4,0.5]	0.0555	0.0403	0.0767	0.0269	0.0188	0.0385
(0.5,0.6]	0.0674	0.0497	0.0917	0.0324	0.0229	0.0458

```
ggplot(interval_df, aes(x = tend, y = `1_cif`)) +
  geom_line() +
  geom_line(aes(x = tend, y = `1_cif_lower`), linetype = "dotted") +
  geom_line(aes(x = tend, y = `1_cif_upper`), linetype = "dotted") +
  geom_line(aes(x = tend, y = `2_cif`), col = 2) +
  geom_line(aes(x = tend, y = `2_cif_lower`), linetype = "dotted", col = 2) +
  geom_line(aes(x = tend, y = `2_cif_upper`), linetype = "dotted", col = 2) +
  ggtitle("Cumulative incidences",
          subtitle = paste("Death from specified cause vs. Death from other",
                           "cause (red) with 95% confidence bands (dotted lines).")) +
  xlab("Time") + ylab("Cumulative incidences")
```

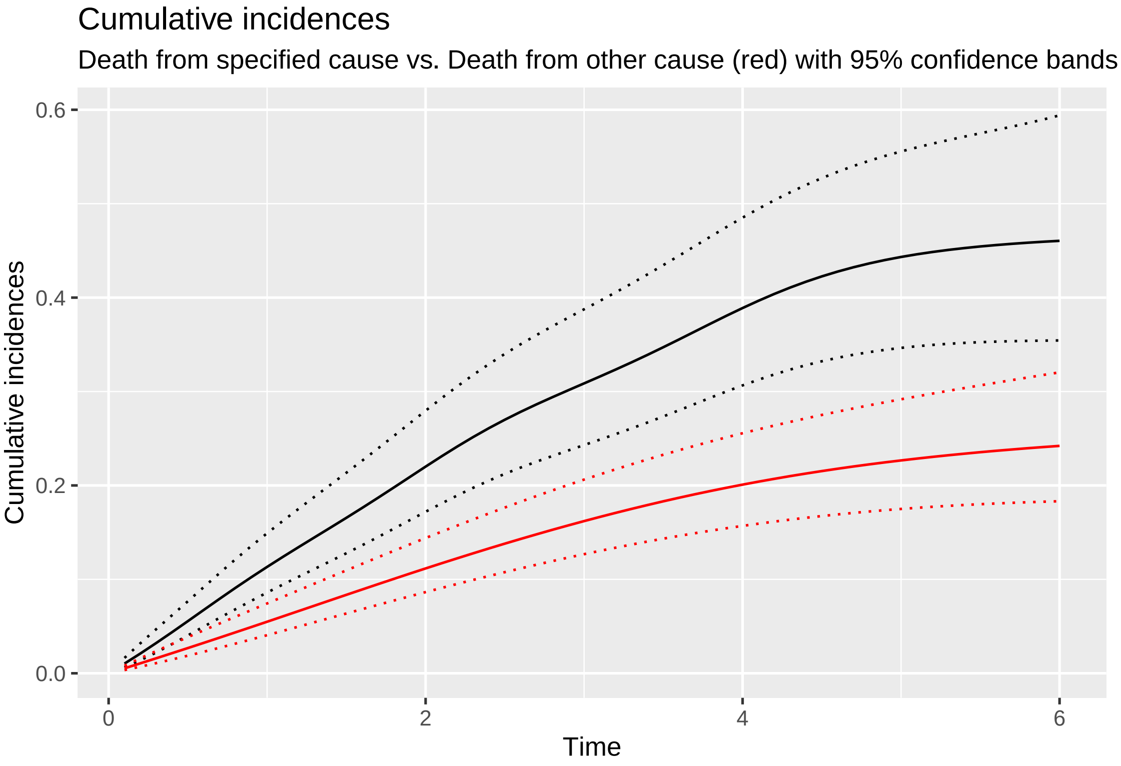

Figure 5.3: CIF estimates using a subdistribution hazards PAM.

5.3 Missing time-dependent covariates

In this section we investigate a problem of PAMMs which already exists in the approach of Fine and Gray (1999), namely missing values of time-dependent covariates.

When modeling subdistribution hazards potentially missing time-dependent covariates may occur. This is because the hypothetical censorship time is larger than the actual one.

In the Appendix we show how `pammtools` processes time-dependent covariates with a focus on potentially missing values. `pammtools` simply carries-on missing values. This means that, from a technical point of view, missing time-dependent covariates are not-problematic. Sometimes, this carrying-on is meaningful. One example for this is if a time-dependent covariate is only reported for mo We show that this means that the missing value problem can be solved outside of `pammtools` **before** transformation to a `ped`. Theoretically, we could complete the whole "missing" time-dependent covariates for the entire follow-up time.

While these insights do not solve the problem described, it makes it model agnostic. The problem has no simplified to an imputation problem for longitudinal data. As an exhaustive analysis of this topic is far beyond the scope of this thesis, we leave it to further research.

Chapter 6

Competing risks PAMMs - an evaluation

The models proposed by this thesis are under certain assumptions equivalent to the Cox proportional hazards model. Competing risks analysis is mainly about controlling the risk set. This means for the same risk and covariate set (and under the presented assumptions) model estimates between a PAMM and the Cox PH model must be equivalent. However, now we want to evaluate how well the models work in practice where the exact equivalence is typically not granted. This chapter does not claim to be an exhaustive simulation study but rather serves as a proof of concepts of the proposed methods.

6.1 Simulating competing risks data

Simulating competing risks data requires some effort and is slightly more complex than single risk simulations. According to Beyersmann et al. (2011) "cause-specific hazards completely determine the stochastic behaviour of the competing risks process". This means when simulating competing risks data, we only care about cause-specific hazards. This also means that we can validate our PAMM models later on by comparing the estimated cause-specific hazards with the true ones. For the simple case of constant hazards with right censoring (left truncation could be easily added) Beyersmann et al. (2011) recommend the following procedure if there are no covariate effects present (Note that we describe the procedure for a single individual.):

1) Simulate failure times using an exponential distribution and all hazards to be investigated. Conveniently, all hazards are additive so that the survival time can be simulated using the sum of all single hazards: $\alpha(t) = \sum_{k=1}^{K} \alpha(t)_k$. The failure time T is derived by a draw from the exponential distribution parametrised with $\lambda = \alpha(t)$.
2) Determine due to which cause the individual failed. This is achieved by a draw from a multinomial distribution. The probabilities at a given event time T for each event (k) are derived by:

$$\pi_k = \frac{\alpha_k(T_i)}{\alpha(T_i)}$$

3) Generate right censoring times (C_i) (and possibly left truncation). This is achieved by drawing a censoring time from an arbitrary (but meaningful) distribution (Note that often an exponential or uniform distribution is chosen).

4) If the simulated survival time T_i of individual i is larger than the corresponding censor time C_i, overwrite the event type by a censoring event and the event time by C_i.

An easy code example which implements the described simulation for **two** competing risk is the following (strongly inspired by Beyersmann et al. (2011)):

```r
set.seed(20200401)
# this simulates the failure - no matter due to which cause.
event_times <- rexp(1000, 1.3) # 1.3 = 0.8 + 0.5
# store hazards / probabilities
h1 <- 0.8
h2 <- 0.5
p1 <- h1 / (h1 + h2)
# this is the binomial experiment which determines the failure cause
cause <- rbinom(1000, size = 1, prob = p1)
# replace 0 by 2
cause <- ifelse(cause == 0, 2, 1)
# simulate censoring via expoential distribution
cens_times <- rexp(1000, 0.2)
# overwrite event time if censor time is smaller
observed_times <- pmin(event_times, cens_times)
# overwrite event type if so with 0 for censoring
cause <- (event_times <= cens_times) * cause

table(cause)
```

0	1	2
161	471	368

Next to this very simple case, we aim to generalise the simulation. We allow covariate effects (which effect the hazard *proportionally* complying with the proportional hazards assumption). This is typically not too hard to facilitate. The hazard will now be different for each individual based on her covariate set. As in the proportional hazards model, one can decompose an individual's hazard into the baseline hazard and multiplicative components. Assume that the categorical feature x_1 affects the hazard for hazard 1 positively (no effect on 2) by an increase positively by, let‘s say, the factor 1.105.

```r
set.seed(20200401)
# simulate x1
x1 <- sample(c(1, 0), 1000, prob = c(0.25, 0.75), replace = TRUE)
# these are the baseline hazards now.
h1 <- rep(0.8, 1000)
h2 <- rep(0.5, 1000)
# increase h1 them if the variable equals to 1
h1 <- h1 * (1.105 * x1)
# simulate the failure times resulting from the hazards
```

```r
event_times <- rexp(1000, h1 + h2)
# compute the individual probabilites.
p1 <- h1 / (h1 + h2)
# this is the binomial experiment which determines the failure cause
cause <- rbinom(1000, size = 1, prob = p1)
# replace 0 by 2
cause <- ifelse(cause == 0, 2, 1)
# simulate censoring via expoential distribution
cens_times <- rexp(1000, 0.2)
# overwrite event time if censor time is smaller
observed_times <- pmin(event_times, cens_times)
# overwrite event type if so with 0 for censoring
cause <- (event_times <= cens_times) * cause

# should be less type 1 events now.
table(cause)
```

0	1	2
274	153	573

This technique can be further generalised to facilitate more complex effects. One example is contributed to this thesis by Andreas Bender.

The code can be accessed here. The function makes use of `pammtools` to simulate **piece-wise exponential** data in the `ped` format in the first place and then use these data to retrieve hazards survival times from a piece-wise exponential distribution. The rest is implemented identically as shown in previously presented code. To work with the function, one needs a `formula`, a `data.frame` object and cut points for the piece-wise exponential distribution. The formula specifies the hazard for each risk. Different risks are separated by a bar. You can either provide a formula without covariate effects:

```r
formula <- ~ log(0.5) | log(0.3)
```

or specify how covariates should affect hazards

```r
formula <- ~ log(0.5) + x1 - x2 | log(0.3) - x1
```

The `data.frame` needs to include covariates only. So we simply simulate some.

```r
data <- cbind.data.frame(x1 = runif (1000, -3, 3), x2 = runif (1000, 0, 6))
```

The simulation itself is only for obtaining simulated event times. Typically, some post-processing (e.g. for censoring) will be necessary. However, censoring is already present in the form of administrative censoring (here at 6).

```r
sim_df <- sim_pexp_cr(formula, data, seq(0, 6, by = 0.25))
head(as.data.frame(sim_df))
```

id	x1	x2	hazard1	hazard2	time	status	type
1	-0.6901	2.4705	0.0212	0.5981	1.3422	1	2
2	-2.0652	3.1688	0.0027	2.3662	0.3232	1	2
3	-1.7015	3.3196	0.0033	1.6446	0.2769	1	2
4	2.2962	1.8216	0.8037	0.0302	1.5014	1	1
5	2.7966	0.7677	3.8026	0.0183	0.3682	1	1
6	1.6660	3.5550	0.0756	0.0567	5.5912	1	1

```
table(sim_df$status)
```

0	1
129	871

```
table(sim_df$type)
```

0	1	2
129	249	622

We see that the `status` only indicates if there was an event, the `type` is more informative. To add right censoring the following code can be employed (Note that this is just a more compact representation of the code from earlier).

```
sim_df <- sim_df %>% mutate(
  cens_time = runif(n(), 0, 4),
  status = if_else(cens_time < time, 0, 1),
  time = pmin(time, cens_time),
  type = status * type)
table(sim_df$type)
```

0	1	2
375	158	467

Essentially, we can "throw away" most columns of the data set. For our survival analysis we only need:

- The covariate set (here `x1` and `x2`).
- The event time (here `time`).
- The status (here a bit confusingly `type`; thus, we rename it).
- And `id` is always helpful, too.

```
sim_df <- sim_df[, c("id", "time", "type", "x1", "x2")]
colnames(sim_df)[3] <- "status"
```

6.2 Recover baselines & CIFs

Bender (2018) show that PAMMs sufficiently work for modeling baseline hazards and a diverse number of effects. As competing risks modeling is mostly about assessing the appropriate risk set, the **main** challenge to our approach is to model the baseline hazards properly. However, later on, we will also investigate the coefficient effects.

To recover baselines we re-simulate the data with a `formula` without covariate effects. Furthermore, we use an exponential distribution for right censoring.

```r
formula <- ~ log(0.5) | log(0.3)
sim_df <- sim_pexp_cr(formula, data, seq(0, 5, by = 0.25)) %>% mutate(
  cens_time = rexp(n(), 0.4),
  status = if_else(cens_time < time, 0, 1),
  time = pmin(time, cens_time),
  type = status * type)
df <- sim_df[, c("id", "time", "type", "x1", "x2")]
colnames(df)[3] <- "status"
```

6.2.1 Cause-specific hazards

For cause-specific hazards, we use the outlined model from the previous chapter. By calling `as_ped_cr_cs()` we make the data to a competing risks `ped` object. However, we also manipulate the event types beforehand for each risk. Competing events are recoded as censoring events for all risks. `pam_cr()` returns the desired model.

```r
ped_cs <- as_ped_cr_cs(data = df, Surv(time, status) ~ ., id = "id",
                       cut = seq(0, max(df$time), 0.05))
pam_cs <- pam_cr(ped_status ~ s(tend), ped = ped_cs, family = "poisson",
                 offset = offset)
```

6.2.1.1 Hazards and cumulative hazards

Hazards directly translate to cumulative hazards being the integral over time of the hazards. The survival function also directly originates from the cumulative hazard function. Thus, comparing cumulative hazards estimates with the ground truth is sufficient for all resulting estimates from the PAMM. We will also compare PAMMs with non-parametric estimators (in this case the Nelson-Aalen estimator). While PAMMs (or PEMs) are in theory capable of recovering these non-parametric estimators, in practice they mostly will not **exactly** do so – especially when modeling a smooth baseline.

We think that a graphical comparison is most meaningful when investigating baseline hazards. Thus, whenever comparing baseline estimates we only make use of graphical analysis.

First, we compute the non-parametric Nelson-Aalen estimates for comparison.

```r
temp01 <- survfit(Surv(time, status == 1) ~ 1, df)
temp02 <- survfit(Surv(time, status == 2) ~ 1, df)
na01 <- cumsum(temp01$n.event / temp01$n.risk)
na02 <- cumsum(temp02$n.event / temp02$n.risk)
na01 <- as.data.frame(cbind(tend = temp01$time, na01 = na01))
na02 <- as.data.frame(cbind(tend = temp02$time, na02 = na02))
```

We use the previously estimated model (`pam_cs`) to predict the baseline hazard explicitly.

```r
interval_df <- int_info(ped_cs[[1]])
head(interval_df)
```

tstart	tend	intlen	intmid	interval
0.00	0.05	0.05	0.025	(0,0.05]
0.05	0.10	0.05	0.075	(0.05,0.1]
0.10	0.15	0.05	0.125	(0.1,0.15]
0.15	0.20	0.05	0.175	(0.15,0.2]
0.20	0.25	0.05	0.225	(0.2,0.25]
0.25	0.30	0.05	0.275	(0.25,0.3]

```
interval_df <- interval_df %>%
  add_cumu_hazard_cr(pam_cs, ci = FALSE)
head(interval_df)
```

tstart	tend	intlen	intmid	interval	1_cumu_hazard	2_cumu_hazard
0.00	0.05	0.05	0.025	(0,0.05]	0.0260	0.0156
0.05	0.10	0.05	0.075	(0.05,0.1]	0.0519	0.0312
0.10	0.15	0.05	0.125	(0.1,0.15]	0.0776	0.0469
0.15	0.20	0.05	0.175	(0.15,0.2]	0.1032	0.0626
0.20	0.25	0.05	0.225	(0.2,0.25]	0.1287	0.0782
0.25	0.30	0.05	0.275	(0.25,0.3]	0.1540	0.0939

Then, we compute the theoretical estimates for a grid of values. (The theoretical cumulative hazard is linear with slopes 0.5 and 0.3.)

```
interval_df$cumhaz1 <- interval_df$tend * 0.5
interval_df$cumhaz2 <- interval_df$tend * 0.3
```

First of all, we compare the estimated cumulative hazard function with the true one.

```
ggplot(interval_df, aes(x = tend, y = `1_cumu_hazard`)) +
  geom_path() +
  geom_path(aes(x = tend, y = `2_cumu_hazard`), col = 2) +
  geom_line(aes(x = tend, y = cumhaz1), linetype = "dotted") +
  geom_line(aes(x = tend, y = cumhaz2), linetype = "dotted", col = 2) +
  ggtitle("Cumulative hazards",
          subtitle = paste("for the competing risks 1 (black) vs. 2 (red).",
                           "Comparison between PAM estimates (solid) \n and",
                           "true values (dotted).")) +
  xlab("Time") + ylab("Cumulative hazards")
```

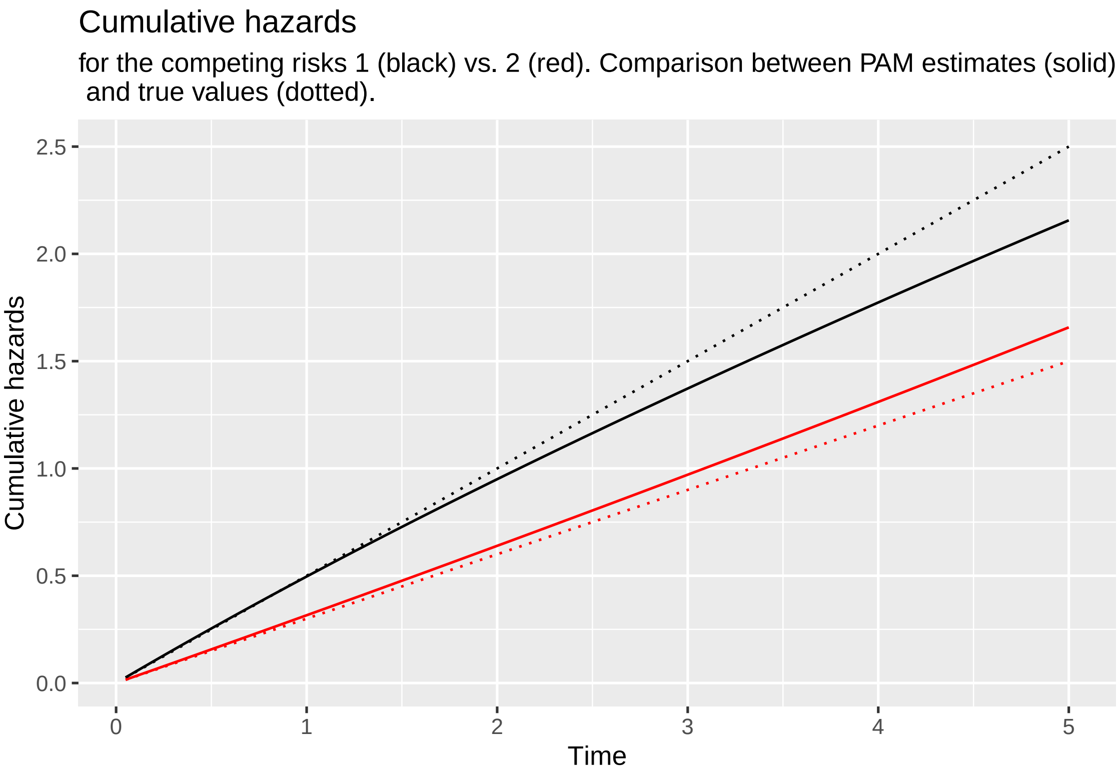

Figure 6.1: Comparison: Cause-specific cumulative hazards - PAM vs. true cumulative hazard rates.

This seems pretty close.

Moreover, we compare PAMM with Nelson Aalen estimates.

```
library(reshape2)
interval_df_melt <-
  melt(interval_df[ , c("tend", "1_cumu_hazard", "2_cumu_hazard")], id = "tend")
na01_melt <- melt(na01, id = "tend")
na02_melt <- melt(na02, id = "tend")
interval_df_melt <- rbind(interval_df_melt, na01_melt, na02_melt)
ggplot(interval_df_melt, aes(x = tend, y = value, colour = variable)) +
  geom_line() + xlab("time") + ylab("Cumulative Hazard") +
  ggtitle("Cumulative hazards",
          subtitle = paste("for the competing risks 1 (black) vs. 2 (red).",
                           "Comparison between PAM (PAM) estimates \n and",
                           "Nelson-Aalen (NA) estimates.")) +
  scale_colour_manual(name = "", values = c("red", "blue", "black", "green"),
                      labels = c("PAM 1", "PAM 2", "NA 1", "NA 2"))
```

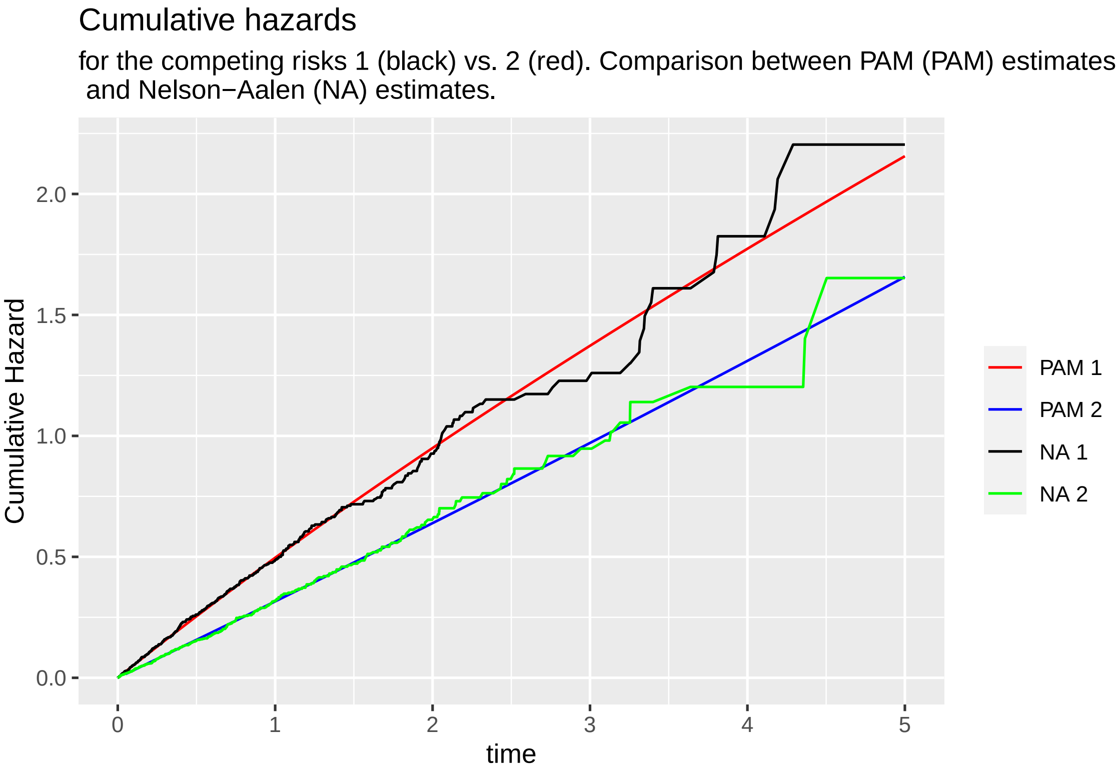

Figure 6.2: Comparison: Cause-specific cumulative hazards estimates - PAM vs. Nelson-Aalen.

We observe that non-parametric and semi-parametric estimates are very close to one another. Very obviously, our PAMM can recover the cause-specific hazards.

From all hazards altogether, one could infer the cumulative incidence function as described in chapter 3. The cumulative incidence function is rather a practical concept, so it makes sense to compare our implementation with the standard implementation from the `cmprsk` package – not with a theoretical measure. Thus, we compare the CIF estimates from the cause-specific hazard PAMM with the estimation by the `cmprsk` package. We also include the confidence intervals in this comparison straight away. The plotting is associated with tedious code which is suppressed in this document.

```
library(cmprsk)
cmprsk_cif <- cuminc(df$time, df$status, cencode = 0)
```

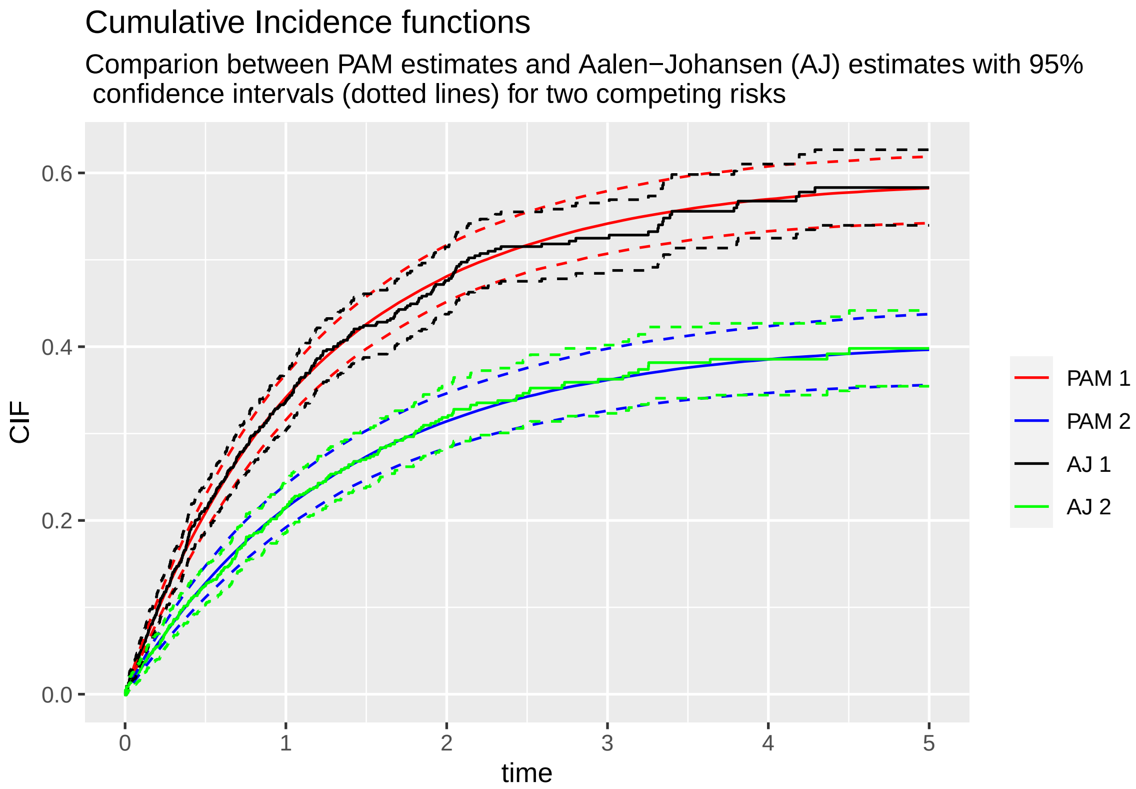

Figure 6.3: Comparison: Cumulative incidence function estimates - Cause-specific PAM vs. Aalen-Johansen.

The smooth semi-parametric estimates replicate the non-parametric `cmprsk` estimates.

6.2.2 Subsdistribution hazards

Subdistribution hazards model the CIF directly and the resulting cumulative hazard function for each risk represents the CIF.

Hence, we will compare the cumulative hazards with the `cmprks` estimates. While Beyersmann et al. (2011) suggest a way to simulate data based on (hypothetical) subdistribution hazards, we decided not to perform this simulation. This is because subdistribution hazards are not directly inferred from the data generating process.

6.2.2.1 No censoring and administrative censoring

First of all, we need to simulate data with administrative censoring only. Recall that we already created such data. We can re-use the code from before and leave out the right censoring. As *no censoring* is a trivial special case of administrative censoring we leave it out at this point.

```
formula_admin <- ~ log(0.25) | log(0.15)
sim_df <- sim_pexp_cr(formula_admin, data, seq(0, 5, by = 0.25))
df_admin <- sim_df[, c("id", "time", "type", "x1", "x2")]
colnames(df_admin)[3] <- "status"
```

This way we introduce administative censoring at $t = 5$. Every single censorship happened at the end of the study.

We use the pipeline from the previous chapter to construct a subdistribution hazards `ped`. We do not need to supply a `max_time` as by default the maximally observed event time is used.

```
ped_sh <- as_ped_cr_sh(data = df_admin, Surv(time, status) ~ ., id = "id",
                       cut = seq(0, max(df_admin$time), 0.05))
pam_sh <- pam_cr(ped_status ~ s(tend), ped = ped_sh, family = "poisson",
                 offset = offset)
```

We predict the CIF from this model and also predict the CIF using `cmprsk`. `cmprsk::cuminc()` actually estimates the CIF for right-censored data. However, administrative censoring means that the weights in the right-censored Fine and Gray model are always one. This is due to the fact the risk set is carried on until the end of the study. Thus, `cmprsk::cuminc()` produces the intended estimates.

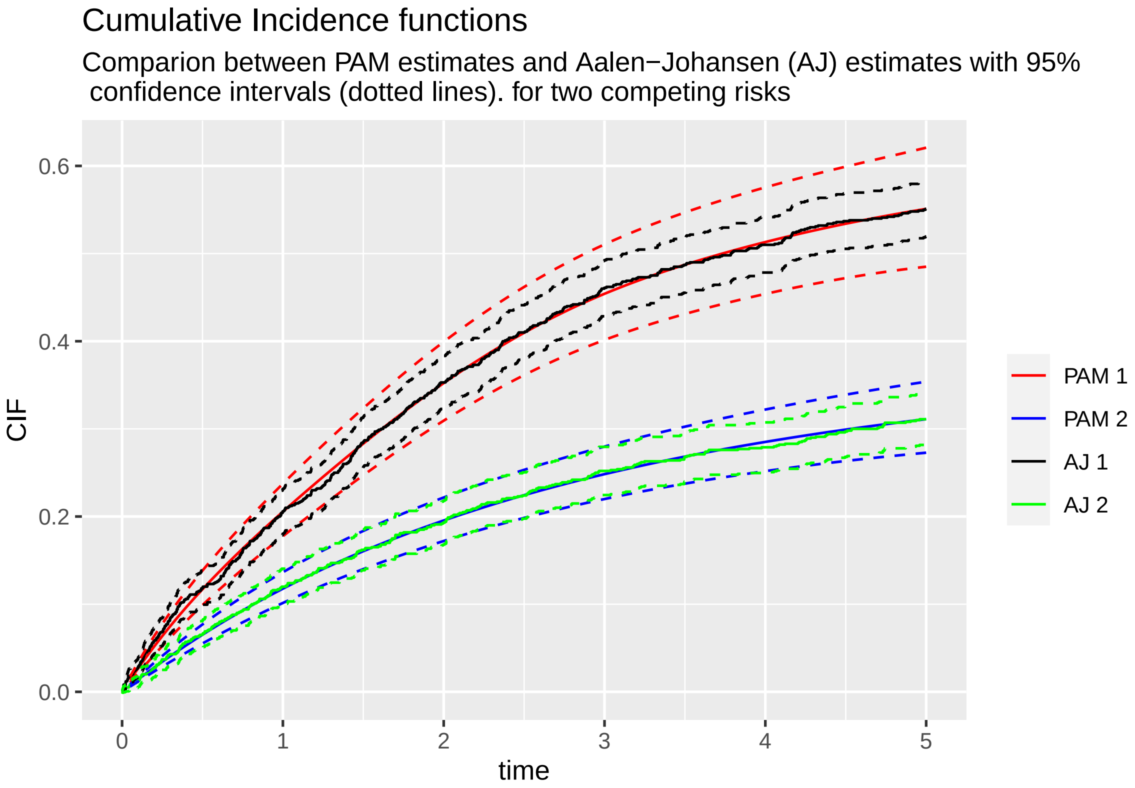

Figure 6.4: Comparison: Cumulative incidence function estimates - Subdistribution PAM vs. Aalen-Johansen I.

The point estimates are recovered – however, the PAMM estimates seem to have higher variance. This is especially apparent for larger values of `tend`. In these areas, there are significantly fewer event times.

Only 7.3 percent of the events happen between `tend = 4` and `tend = 5`. The semi-parametric estimation in this area is naturally subject to higher uncertainty. The variance of the non-parametric

CIF is based on the Nelson-Aalen estimator. Its variance is (from Beyersmann et al. (2011)):

$$\hat{\sigma}^2(t) = \sum_{t_j \le t} \frac{(r_j - d_j)d_j}{(r_j - 1)r_j^2}$$

where r_j is the magnitude of the risk set just before $t = t_j$ and d_j the number of failed individuals until then. C.p. if the risk set increases, in the limit the denominator increases stronger than the numerator. This means that for a large risk set (which is typically the case for administrative censoring) the variance will tend to be small. Furthermore, if there are only a few events, the variance does not increase much. Thus, the difference between larger values of `tend` can be entirely reasoned by a non-parametric or semi-parametric estimation procedure.

In the limit, the variances should be even more similar, though. If we re-run the analysis with more data (5000 simulated observations) the absolute difference decreases but persists.

```r
data2 <- cbind.data.frame(x1 = runif (5000, -3, 3), x2 = runif (5000, 0, 6))

sim_df2 <- sim_pexp_cr(formula_admin, data2, seq(0, 5, by = 0.25))
df_admin <- sim_df2[, c("id", "time", "type", "x1", "x2")]
colnames(df_admin)[3] <- "status"

ped_sh <- as_ped_cr_sh(data = df_admin, Surv(time, status) ~ ., id = "id",
                       cut = seq(0, max(df_admin$time), 0.05))
pam_sh <- pam_cr(ped_status ~ s(tend), ped = ped_sh, family = "poisson",
                 offset = offset)
```

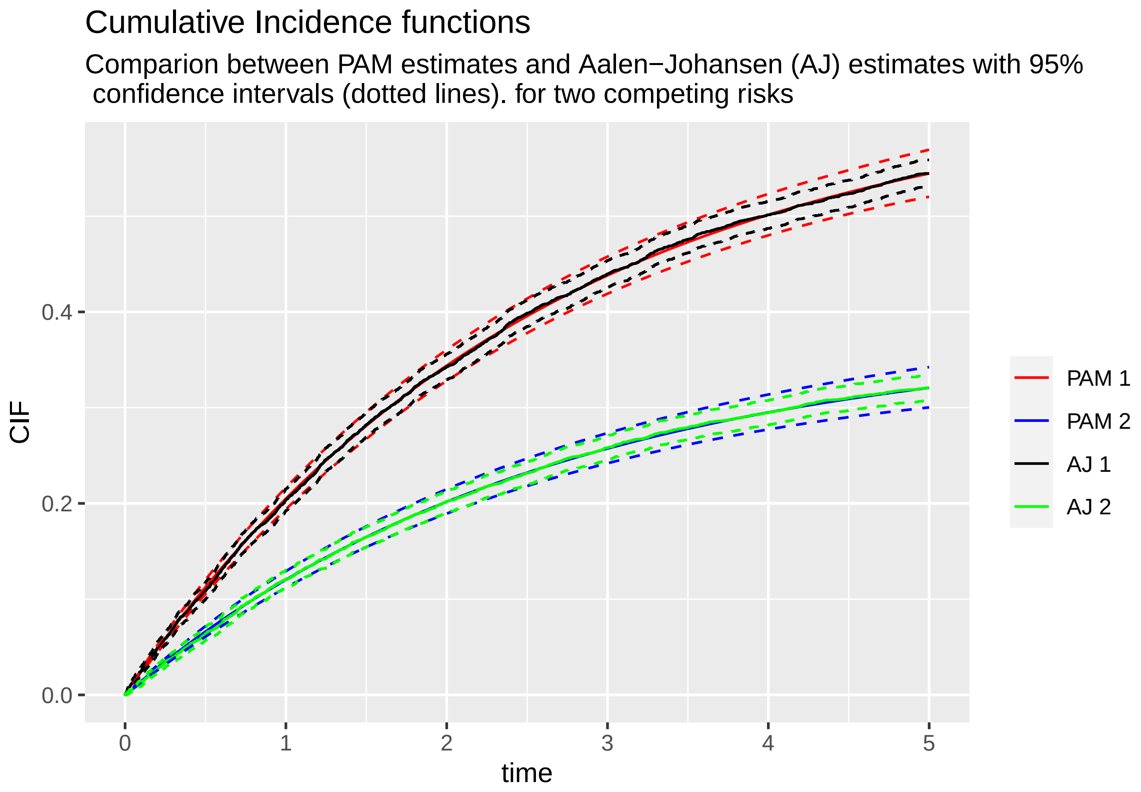

Figure 6.5: Comparison: Cumulative incidence function estimates - Subdistribution PAM vs. Aalen-Johansen II.

Note that we think that this property is rather a strength of PAMMs. Furthermore, it can be easily explained by the nature of the semi-parametric estimation procedure. The PAMM estimation accounts for the complicated data situation at later points in time (fewer events paired with many clustered events at the end of the study) pretty well.

In total, we still can confirm that the PAMM works analogously to the non-parametric estimation, though. PAMMs can model baseline hazards – and resulting CIFs for administratively censored data properly.

6.2.2.2 Right censored data

For right-censored data, we use `as_ped_cr_cens()` in the preprocessing which simulates appropriate censorship times from the underlying censorship distribution. We model this distribution as piece-wise exponential distribution via a PAMM.

For our simulation, we use the initially simulated data for the cause-specific hazards model.

```
ped_cens <- as_ped_cr_cens(data = df, Surv(time, status) ~ ., id = "id",
                           cut = seq(0, max(df$time), 0.05))
pam_cens <- pam_cr(ped_status ~ s(tend), ped = ped_cens, family = "poisson",
                   offset = offset)
```

Again, we compare the PAMM CIF estimates with the `cmprks` estimates graphically.

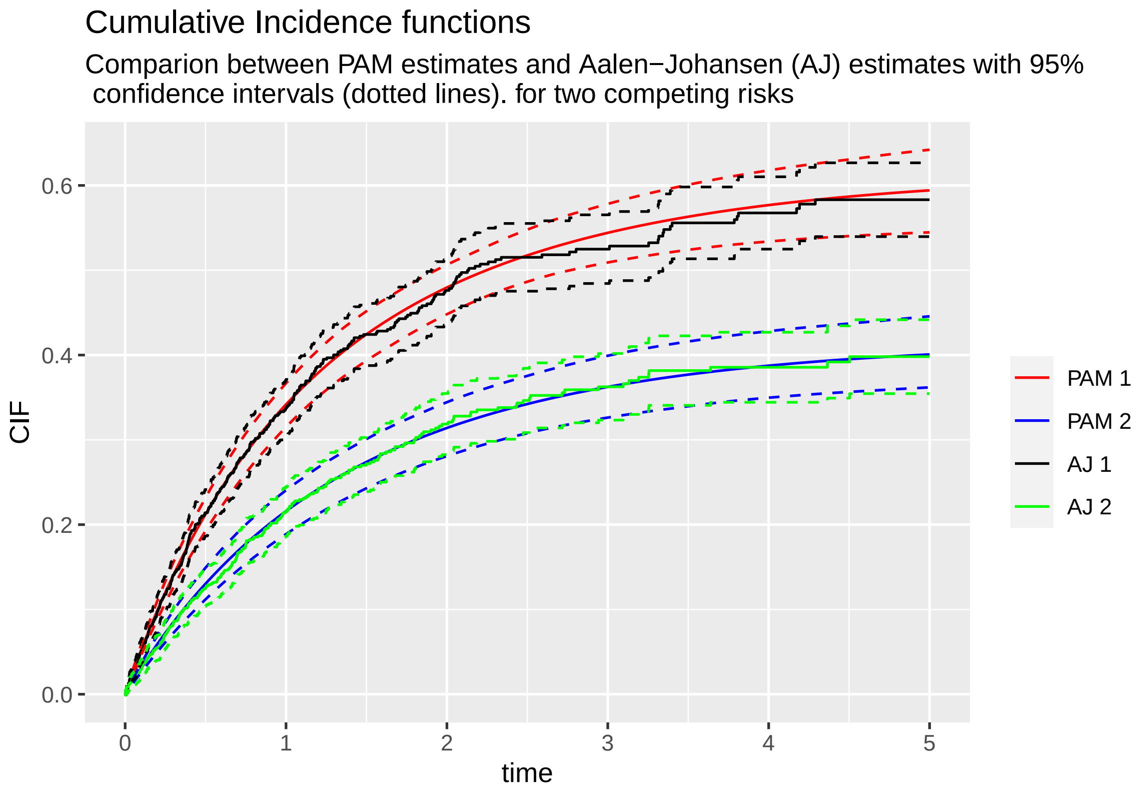

Figure 6.6: Comparison: Cumulative incidence function estimates - Subdistribution PAM vs. Aalen-Johansen III.

Like before we see very good matches between the two approaches. We empirically confirm that the simulation approach can replace the weighting used by Fine and Gray (1999).

However, in this context, the underlying censorship distribution was an exponential distribution. We explicitly model a piece-wise exponential distribution for the censorship distribution. If we use data with a censorship distribution which does not correspond to the piece-wise exponential distribution, we might find that the two approaches will be different. This is because the approach by Fine and Gray (1999) remains non-parametrical while our approach rests upon the parametric assumption of a (piece-wise) exponential distribution. A non-parametric baseline is agnostic to the underlying actual distribution.

```
sim_df <- sim_pexp_cr(formula, data, seq(0, 6, by = 0.25))
sim_df <- sim_df %>% mutate(
  cens_time = runif(n(), 0, 8),
  status = if_else(cens_time < time, 0, 1),
  time = pmin(time, cens_time),
  type = status * type)
df <- sim_df[, c("id", "time", "type", "x1", "x2")]
colnames(df)[3] <- "status"

ped_cens <- as_ped_cr_cens(data = df, Surv(time, status) ~ ., id = "id",
                           cut = seq(0, max(df$time), 0.05))
pam_cens <- pam_cr(ped_status ~ s(tend), ped = ped_cens, family = "poisson",
                    offset = offset)
```

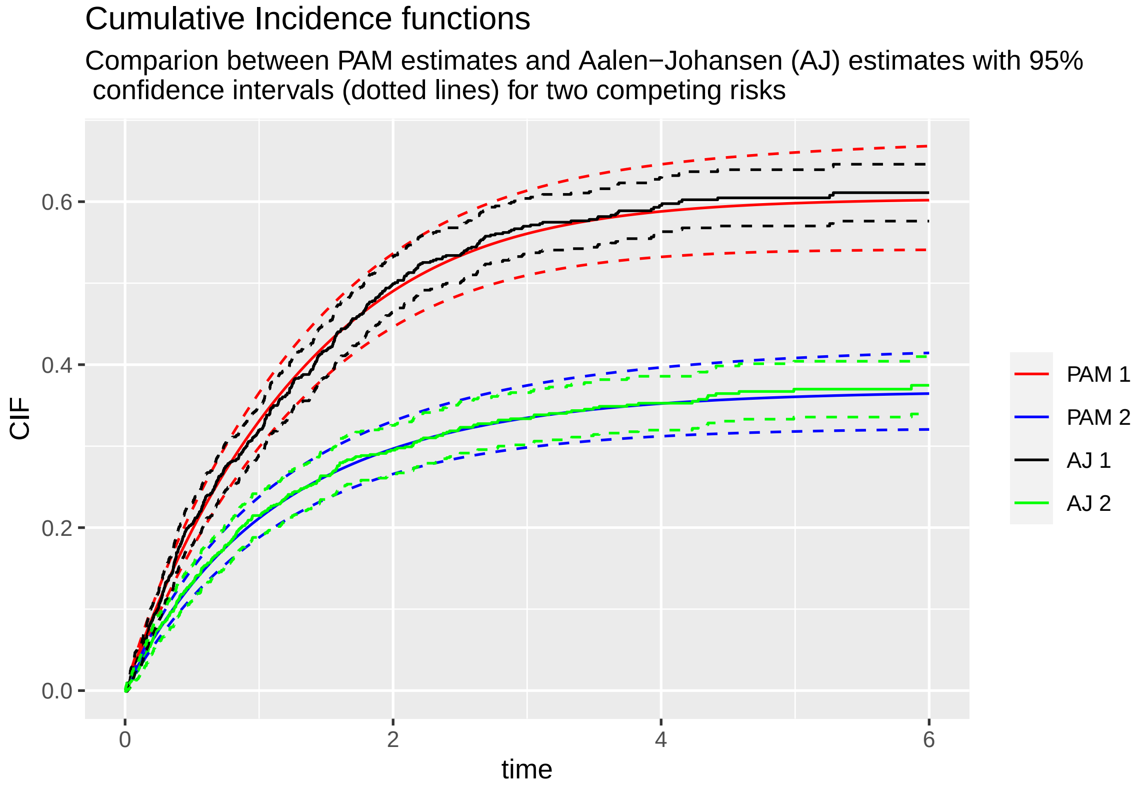

Figure 6.7: Comparison: Cumulative incidence function estimates - Subdistribution PAM vs. Aalen-Johansen IV.

We observe that `cmprsk` and our PAMM now tend to predict different CIFs (this implies different subdistribution baseline hazards) for later points in time. The `cmprsk` estimation remains stable and in this setting, it can replicate the uniform distribution better – non-parametric confidence intervals also seem to be smaller. The current implementation of the subdistribution PAMM (for incomplete data) demands that the censorship distribution is sufficiently well explained by a piece-wise exponential distribution.

However, we aim to emphasise that it is the most natural assumption within survival analysis that survival times originate from a(n) (piece-wise) exponential distribution. Thus, we think a transfer of this assumption to censorship times is not extremely critical. For a more general translation of this assumption, there is certainly more research necessary, though. Still, the current implementation is a well-working vehicle for the richness of applications.

6.3 Recover coefficients

For the remaining sanity checks, we assume that censorship also follows an exponential distribution. This makes sure that we can focus on coefficient effects. However, we will illustrate the impact of this misspecification and analyse its impact on covariates in this simple setting as well.

6.3.1 Static covariates

We start away with the simplest kind of effect: static covariates.

Thus, we now simulate two different data sets with covariate effects. One with exponential censoring and one with uniform censoring. Additionally, we simulate a data set with administrative censoring only.

```r
formula <- ~ log(0.3) + x1 + 0.2 * x2 | log(11) - 4 * x2
sim_df <- sim_pexp_cr(formula, data, seq(0, 6, by = 0.25)) %>% mutate(
  cens_time = rexp(n(), 0.2),
  status = if_else(cens_time < time, 0, 1),
  time = pmin(time, cens_time),
  type = status * type)
sim_df <- sim_df[, c("id", "time", "type", "x1", "x2")]
colnames(sim_df)[3] <- "status"
table(sim_df$status)
```

0	1	2
311	572	117

```r
sim_df_unif <- sim_pexp_cr(formula, data, seq(0, 6, by = 0.25)) %>% mutate(
  cens_time = runif(n(), 0, 6),
  status = if_else(cens_time < time, 0, 1),
  time = pmin(time, cens_time),
  type = status * type)
sim_df_unif <- sim_df_unif[, c("id", "time", "type", "x1", "x2")]
colnames(sim_df_unif)[3] <- "status"
table(sim_df_unif$status)
```

0	1	2
330	564	106

```r
sim_df_admin <- sim_pexp_cr(formula, data, seq(0, 6, by = 0.25))
sim_df_admin <- sim_df_admin[, c("id", "time", "type", "x1", "x2")]
colnames(sim_df_admin)[3] <- "status"
table(sim_df_admin$status)
```

0	1	2
188	681	131

6.3.1.1 Cause-specific hazards

For cause-specific hazards, we employ our usual workflow and compare the model output(s) of the Cox proportional hazards model(s) (in R) with the model output of our PAMM.

```r
library(survival)
df <- sim_df
cox_df <- df
cox_df$SurvObj1 <- with(cox_df, Surv(time, status == 1))
cox_df$SurvObj2 <- with(cox_df, Surv(time, status == 2))
```

```r
cox_model_cs1 <- coxph(SurvObj1 ~ x1 + x2, data = cox_df)
cox_model_cs2 <- coxph(SurvObj2 ~ x1 + x2, data = cox_df)
ped_cs <- as_ped_cr_cs(data = df, Surv(time, status) ~ ., id = "id",
                       cut = seq(0, max(df$time), 0.05))
pam_cs <- pam_cr(ped_status ~ s(tend) + x1 + x2, ped = ped_cs,
                 family = "poisson", offset = offset)
cox_coeffs <- rbind(cox_model_cs1$coefficients, cox_model_cs2$coefficients)
rownames(cox_coeffs) <- c("cox1", "cox2")
pamm_coeffs <- rbind(pam_cs$`1`$coefficients[c("x1", "x2")],
                     pam_cs$`2`$coefficients[c("x1", "x2")])
rownames(pamm_coeffs) <- c("pamm1", "pamm2")
rbind(cox_coeffs, pamm_coeffs)
```

	x1	x2
cox1	1.0918	0.1738
cox2	0.0156	-3.9815
pamm1	1.0841	0.1731
pamm2	0.0142	-4.0253

The estimated coefficients are very similar. In this simple setting, the PAMM recovers all cox model coefficients. Additionally, the true coefficients seem to be pretty well recovered.

Also the standard errors appear to be very similar., e.g. consider:

```r
data.frame(se_pam1 = summary(pam_cs$`1`)$se["x1"],
           se_cox1 = summary(cox_model_cs1)$coefficients["x1", "se(coef)"])
```

	se_pam1	se_cox1
x1	0.0409	0.0417

6.3.1.2 Subsdistribution hazards

Subdistribution hazards do not – unlike the cause-specific hazards – describe the data generating process. Thus, for subdistribution hazards, the comparison with true population parameters is not feasible. We only compare the Fine and Gray Cox model estimates with PAMM estimates.

6.3.1.2.1 No censoring and administrative censoring

First, we investigate the model for no censoring or (less trivially) administrative censoring. Administrative censoring can be facilitated by `cmprsk::crr()`. The function implements the more general case of right censoring, though. However, as before this model for administrative censoring is only a special case of the right-censored one.

```r
df <- sim_df_admin
cox_model_sh1 <- crr(df$time, df$status, cox_df[, c("x1", "x2")], failcode = 1)
cox_model_sh2 <- crr(df$time, df$status, cox_df[, c("x1", "x2")], failcode = 2)
ped_sh <- as_ped_cr_sh(data = df, Surv(time, status) ~ ., id = "id",
                       cut = seq(0, max(df$time), 0.05), max_time = 6L)
pam_sh <- pam_cr(ped_status ~ s(tend) + x1 + x2, ped = ped_sh,
                 family = "poisson", offset = offset)
cox_coeffs <- rbind(cox_model_sh1$coef, cox_model_sh2$coef)
rownames(cox_coeffs) <- c("cox1", "cox2")
pamm_coeffs <- rbind(pam_sh$`1`$coefficients[c("x1", "x2")],
                     pam_sh$`2`$coefficients[c("x1", "x2")])
rownames(pamm_coeffs) <- c("pamm1", "pamm2")
rbind(cox_coeffs, pamm_coeffs)
```

	x1	x2
cox1	0.7651	0.3863
cox2	-0.3531	-3.2459
pamm1	0.7729	0.3897
pamm2	-0.3573	-3.2781

Also in this scenario point estimates are very similar. The problem with the estimates is, though, that they are not as much related to the original risk process anymore as the cause-specific ones.

6.3.1.2.2 Right censored data

For right-censored data, we can apply almost the identical code as before. We only use different data and `as_ped_cr_cens()`.

```
df <- sim_df
cox_df <- df
cox_model_cens1 <- crr(cox_df$time, cox_df$status, cox_df[, c("x1", "x2")],
                       failcode = 1)
cox_model_cens2 <- crr(cox_df$time, cox_df$status, cox_df[, c("x1", "x2")],
                       failcode = 2)
ped_cens <- as_ped_cr_cens(data = df, Surv(time, status) ~ ., id = "id",
                           cut = seq(0, max(df$time), 0.05))
pam_cens <- pam_cr(ped_status ~ s(tend) + x1 + x2, ped = ped_cens,
                   family = "poisson", offset = offset)
cox_coeffs <- rbind(cox_model_cens1$coef, cox_model_cens2$coef)
rownames(cox_coeffs) <- c("cox1", "cox2")
pamm_coeffs <- rbind(pam_cens$`1`$coefficients[c("x1", "x2")],
                     pam_cens$`2`$coefficients[c("x1", "x2")])
rownames(pamm_coeffs) <- c("pamm1", "pamm2")
rbind(cox_coeffs, pamm_coeffs)
```

	x1	x2
cox1	0.8818	0.3642
cox2	-0.3208	-3.1570
pamm1	0.8982	0.3645
pamm2	-0.3269	-3.2984

Also in this scenario, point estimates seem reasonably related.

Now we rerun the analysis, however with uniform censoring – deliberately violating one assumption of our PAMM.

```
df <- sim_df_unif
cox_df <- df
cox_model_cens1 <- crr(cox_df$time, cox_df$status, cox_df[, c("x1", "x2")],
                       failcode = 1)
cox_model_cens2 <- crr(cox_df$time, cox_df$status, cox_df[, c("x1", "x2")],
                       failcode = 2)
ped_cens <- as_ped_cr_cens(data = df, Surv(time, status) ~ ., id = "id",
                           cut = seq(0, max(df$time), 0.05))
pam_cens <- pam_cr(ped_status ~ s(tend) + x1 + x2, ped = ped_cens,
                   family = "poisson", offset = offset)
cox_coeffs <- rbind(cox_model_cens1$coef, cox_model_cens2$coef)
rownames(cox_coeffs) <- c("cox1", "cox2")
pamm_coeffs <- rbind(pam_cens$`1`$coefficients[c("x1", "x2")],
                     pam_cens$`2`$coefficients[c("x1", "x2")])
rownames(pamm_coeffs) <- c("pamm1", "pamm2")
rbind(cox_coeffs, pamm_coeffs)
```

	x1	x2
cox1	0.8638	0.3284
cox2	-0.3380	-3.1134
pamm1	0.8548	0.3431
pamm2	-0.3432	-3.1080

Not like for the baseline hazard, the misspecification does not seem to create any practical difference in parameter estimates.

6.3.2 Time-dependent covariates & more complex covariates

One of the main motivations of PAMMs is the straight forward modeling of more complex effects. While, for the cause-specific hazards model, we may compare our model estimates very well (the `survival` package has nice implementations of the extended cox model), it is much harder to compare subdistribution model estimates.

Comparing cause-specific hazards is rather trivial and has been done by Bender and Scheipl (2018) already. Furthermore, Bender et al. (2018b) validate that even highly complex ELRAs can be recovered in an extensive simulation study for the subdistribution hazards model in the presence of administrative censoring. While an equivalent approach for the newly invented right-censored subdistribution PAMM would be appropriate to study, it is far beyond the scope of this thesis and left for future research.

6.4 Summary and discussion

We showed that PAMMs are capable of modeling rather simple associations as well as the Cox model. Essentially, these can be seen as proof of concepts. The models work in the way we expect them to. The semiparametric estimation of the PAMMs works in general for all presented settings.

More complex investigations are beyond the scope of this thesis, however, will be necessary to finally approve the use of the models presented in this thesis.

Chapter 7

Data analysis

This chapter aims to re-assess the ELR associations estimated by Bender et al. (2018b). However, we use an updated version of the data featuring more than 16000 patients. The new data also covers the years 2013 and 2014.

Bender et al. (2018b) focus on the effects of nutrition intake on the cumulative incidence function (by assuming a hypothetical censoring at the end of the study for discharged individuals). We study the risk process in greater depth by an exhaustive competing risks analysis. First of all, we investigate the competing event "discharge from hospital" explicitly. This helps in understanding the risk process as a whole. Furthermore, we also investigate cause-specific hazards. As outlined in chapter 4, cause-specific hazards are used if causal effects are to be investigated. We agree with Bender et al. (2018b) that causal inference in a confounded observational design like their's is very hard. Still, the modeling of cause-specific hazards is the standard approach if covariate effects are intended to be interpreted causally. We analyse the ELRA by a subdistribution and a cause-specific hazards model and investigate the model inferences *vis-à-vis*, following Beyersmann et al. (2011) . Next to the impact of caloric intake, we also analyse the effect of protein intake on survival and discharge.

7.1 Motivation

Next to the motivation of Bender et al. (2018b) to model ELRA for the given data set which is mainly a technical (see chapter 3), we outline why the association is relevant to study from a clinical point of view. There is no doubt that artificial nutrition is an important tool for intensive medicine to save lives. Nevertheless, it is not clear to what extent artificial nutrition is beneficial. Effects may be lagged and varying strongly over time. Also, confounding is an eminent problem. These complexities lead to strong disagreement on the optimal administration of artificial nutrition. For example, Patel et al. (2017) outlines that there are strong differences in the guidelines for American and Canadian practitioners. Nevertheless, there is great consent that malnutrition leads to increased hospital mortality (Cederholm et al. (1995)). In the long-run individuals should, of course, receive the prescribed amount of food. In the short-run, a critical illness is a hypercatabolic state in which the human body develops specific inflammatory and metabolic responses (Sharma et al. (2019)). This state may require a differential nutrition administration. As outlined in Hartl et al. (2019) there is a richness of opinions on what a proper diet should be on the ICU ranging from hypocaloric diets to high-caloric diets. Next to calories only (as performed in Bender et al. (2018b)) protein intake may be also interesting to study. This is because proteins play a substantial role in the physiological healing process. In particular, Hoffer (2016) outlines that the supply of protein plays a more substantial role in the healing process than calories and should be focused on in future research.

It is highly relevant to study the effects of nutrition on hospital survival for obvious reasons: Heyland et al. (2011) explain that in the long-run an energy deficit is associated with decreased medical outcomes.

In this analysis, we also study the impact of nutrition on hospital discharge time. A decreased admission time in the ICU is associated with lower medical costs, extended well-being of the patients, and increased intensive medical capacities.

7.2 Data

In our analysis we use the same data source as Bender et al. (2018b): The data covers patients who were consecutively intubated, over the age of 18 and mechanically ventilated within the first 48 h in the ICU. The patients had to remain at least 72 h in the ICU to be considered in the study. For each patient's relevant information on the patient status (sex, weight, height, etc.), the health status (Apache II score, admission category, diagnosis category, etc.) have been collected on admission. Soon after admission nutrition goals (proteins and calories) have been determined by a dietitian or physician. During the time in the ICU, the nutrition protocols are tracked daily for maximally 12 days. The study distinguishes between oral intake and both, enteral and parenteral nutrition.

In total there are 16047 patients from 774 distinct ICUs. Around 25% of these patients died in the course of the study.

The following sections deal with the preprocessing of the data and some descriptive analyses.

7.2.1 Preprocessing

Most of the preprocessing is very tedious and hence only described verbally. The `R` code associated with it can be accessed in the repository of this thesis.

We describe the preprocessing which we conducted in great detail in the following sections. Afterward, we validate the data by comparing it to the previously used data. For a description of the tedious data cleaning procedures, refer to the Appendix.

While some of the variables from the database can be directly used (such as information on the Apache II score), many relevant features need to be constructed. These features are presented in this section.

In contrast to Bender et al. (2018b), we investigate both risks and construct three event times: `event`, `Surv0To60`, `Disc0To60`. Event times larger than 61 are censored to 61. These event times are derived via the time difference of the admission time (into the hospital) and death of the patient or hospital discharge, respectively.

The diagnosis category (`DiagID2`) needs to be derived out of a very long codebook of diagnosis codes. We end up with 9 categories: “Cardio-Vascular”, “Respiratory”, “Gastrointestinal”, “Neurologic”, “Sepsis”, “Orthopedic/Trauma”, “Metabolic”, “Renal” and “Other”.

The frequency of mechanical ventilation between days two and four (`MV2_4`) needs to be derived via two steps: The first step is the more complex derivation of a dummy indicating for each study day whether the patient was in mechanical ventilation. The second step is the grouping and summation over the dummies (ones and zeros) over days 2 until 4. The same is true for the oral intake (`OralIntake2_4`) parenteral nutrition (`PN2_4`) and propofol (`Propofol2_4`) administration between days 2 and 4.

The nutritional intake of the patient is subject to the most extensive preprocessing which will be explained in the following.

Nutritional intake is recorded daily. This discrete measurement is not a big problem but interferes at the edges (i.e. admission and discharge/death) with the more granular time measurements of the event time. For example, if a patient’s admission to ICU was at 11 pm, he will most likely not have received any food on day one. Thus, Bender et al. (2018b) suggest to rather operate on complete **calendar** days.

Recall that Bender et al. (2018b) used the following three nutrition categories:

- C I: less or equal than 30% of prescribed calories.
- C II: between 30% (more than) and 70% (less or equal) of prescribed calories **or** less or equal than 30% with additional oral intake.
- C III: more than 70% of prescribed calories **or** between 30% and 70% with additional oral intake.

We use the same classification in our analysis.

For protein, we also use three categories. These categories are based on the following scheme:

- C I: less than 0.8 g protein per Kg body weight.
- C II: between (more or equal) 0.8 g and (less than) 1.2 g protein per Kg body weight.
- C III: more (or equal) than 1.2 g protein per Kg body weight.

For protein, additional oral intake is not assumed to move the individual one category up. This is because we want to study the isolated effect of artificial protein supply.

Both classifications have been recommended by the medical advisor of this thesis, Wolfgang Hartl. For caloric intake the classification is heuristically meaningful. Furthermore, similar classifications are also suggested by related literature (e.g. Krishnan et al. (2003) or Heyland et al. (2011)) For protein, typically an intake of more than 1.2 g per Kg body weight is considered a high protein intake (e.g. Singer et al. (2019) or Weijs et al. (2019)). According to Hoffer (2016) normal prescriptions of protein intake resemble to 0.8 g per Kg body weight. They also outline that a common suggestion for critical-ill patients is 1.5 g per day. With a protein supply of at least 1.2 g per day per Kg body weight, 80 percent of the prescribed target of 1.5 g would be achieved. Thus, the protein categories can be understood as:

- C I: Subnormal
- C II: Normal

- C III: Target

These covariates are represented – like in Bender et al. (2018b) – by two dummy covariates. One dummy

A substantial modeling assumption put into the preprocessing is how to deal with incomplete nutrition protocols. Bender et al. (2018b) assume that patients discharged from the ICU were presumably healthier than the remaining ones and thus they assumed a C III for the not observed study days. However, this assumption is very strong and may cause unintentional confounding as causality is reverted in this scenario. As outlined in chapter 5, one way to deal with incomplete time-dependent covariates is by carrying them on. For cumulative effects that means that the last observed total exposure is carried on. In our scenario, this has an undesired implication: Like Bender et al. (2018b), we aim to describe the nutritional effect of two categorical indicators for the two upper categories. A carry-on method would implicitly assume a C I intake for all incomplete days. Following Bender et al. (2018b), we think a C III assumption fits the situation the best.

While discharge from ICU perfectly explains the reason for the missingness of the data, there are also incomplete protocols for other (non-explainable) reasons. A priori it is not clear how to deal with this sort of missing values.

As discussed in chapter 5, this problem is a longitudinal data imputation problem and beyond the scope of this thesis. However, we try out three different assumptions in our analysis: We deal with all incomplete data identical. In the main specification, we assume for missing data (for whatever reason) a C III intake for calories and protein.

7.2.2 Mapping with the previous data

Due to some new conventions in the data (such as a new `id` administration), it is not possible to directly identify identical observations in our data and the data of Bender et al. (2018b). The detailed results from the mapping can be studied in the Appendix. We only report the results here. Apart from (**slightly**) differently engineered (around 1000 patient-days), we find the data to be consistent. However, more than 20000 patient days could not be mapped. While the remainder of patient days seem to be very consistent, a complete checkup on this would be beneficial.

7.2.3 Descriptive analysis

To get a feeling for the kind of data that is modeled, we start by reporting the general summary statistics. For these summary statistics, it is necessary to distinguish between the two data sets very carefully. While the features of the `patient` data set give information on patients, the `daily` data set gives information on patient days. There are maximally 12 study days for each individual available in which the administration of artificial nutrition is studied.

```r
source("code/helpers.R")
patient <- readRDS("data/patient.Rds")
daily <- readRDS("data/daily.Rds")
merged <- readRDS("data/mergedAndCleanedData.Rds")
```

In total, we gather relevant information on the following variables:

- event: The number of days until any event happened. For each type of event (death in hospital and discharge) the survival time is tracked in Surv0To60 and Disc0To60, respectively.
- PatientDied: An indicator of the type of event (1 death in hospital).
- PatientDischarged: An indicator of the type of event (1 discharge).
- Year: the year of admission for each patient.
- DiagID2: the diagnosis category.
- AdmCatID: the admission category.
- Gender: a patient's sex.
- ApacheIIScore: a patient's Apache II score on admission day.
- BMI: the reported BMI on admission day.

Bender et al. (2018b) derive caloric and protein adequacy (actual intake / prescribed calories) in percent (caloriesPercentage and proteinAdjustedPercentage). They do so by taking the nutrition intake and adjusting it by the prescribed amount (calories) or the bodyweight (protein). As these values are expected to be a noisy error, we construct (as outlined in the previous section) a categorical feature calCat which separates the data into three intervals (less or equal than 30% of prescribed intake, more than 70% prescribed intake and one category for in between). calCat is stored in two separate dummy variables for C II (calCat2) and C III (calCat3). The same is true for the protein intake (proteinCat2, proteinCat3).

This information is paired with an ID, namely CombinedID and an ID for all 774 distinct ICUs (CombinedicuID). Furthermore, the data contains the following relevant features, which are, however, not examined descriptively:

- Propofol2_4: the number of propofol administrations between the second and fourth days in ICU.
- OralIntake2_4: the number of oral intakes between the second and fourth days in ICU.
- PN2_4: the number of parenteral artificial nutrition administrations between the second and fourth days in ICU.
- EN2_4: the number of enteral artificial nutrition administrations between the second and fourth days in ICU.
- MV2_4: the number of mechanical ventilation between the second and fourth days in ICU.

The following tables examine relevant patient data. The rest of the section outlines will describe these tables and further investigate interesting potential insights.

Table 7.1: Key summaries of categorical patient data (numbers reflect individual patients).

Variable	Technical name	Categories	Obs.	Percent	Mortality
Year	`year`	2007	2136	13.31%	27.77%
		2008	2156	13.43%	26.99%
		2009	2339	14.57%	24.11%
		2011	3425	21.34%	23.56%
		2013	3185	19.85%	25.15%
		2014	2806	17.49%	24.41%
Sex	`gender`	Male	9813	61.15%	24.70%
		Female	6234	38.85%	25.79%
Admission Category	`AdmCatId`	Medical	10378	64.67%	28.18%
		Surgical Elective	1641	10.23%	19.74%
		Emergency	4028	25.10%	19.44%
Diagnosis Category	`DiagID2`	Respiratory	3635	22.65%	29.68%
		Cardio-Vascular	2234	13.92%	27.15%
		Gastrointestinal	2003	12.48%	25.11%
		Neurologic	2095	13.06%	23.24%
		Sepsis	1711	10.66%	32.20%
		Orthopedic	1927	12.01%	13.65%
		Metabolic	299	1.86%	14.38%
		Renal	114	<1%	34.21%
		Other	2029	12.6%	24.44%

Table 7.2: Key summaries of numeric patient data.

	Age (`Age`)	Apache II Score (`ApacheIIscore`)	BMI (`BMI`)
Min	18.0	0.0	13.1
1. Quartile	48.0	17.0	22.6
Median	62.0	22.0	25.7
Mean	59.9	22.3	27.3
3. Quartile	73.0	28.0	30.1
Max	102.0	145.0	108.5

Table 7.3: Key summaries on nutrition protocols (numbers reflect patient days).

	Category I	Category II	Category III
Calories	31075 (17.60%)	38502 (21.81%)	106936 (60.58%)
Protein	74399 (42.14%)	48247 (28.68%)	53867 (42.15%)

One can see that the new study years 2013 and 2014 such as the last study year of the previous analysis of Bender et al. (2018b), 2011, are overrepresented compared to earlier study years. 2013 and 2014 account for 5991 observations or around 37 percent of the data used in the analysis. Lethality with circa 25 percent has increased compared to the data of Bender et al. (2018b) with 20 percent. However, Bender et al. (2018b) only regard survival in the first 30 days. When only interested in the first 30 days the mortality drops to 21 percent.

We separately investigate survival and discharge waiting time.

Patients tend to die sooner than they are discharged. Both distributions can be reasonably well explained by an exponential distribution:

```
patient_alive <- patient[patient$PatientDied == 1, ]
ggplot(patient_alive, aes(x = event)) + geom_histogram() +
  xlab("Survival time (days)") + ylab("Frequency") +
  ggtitle("Survival times of patients who died")
```

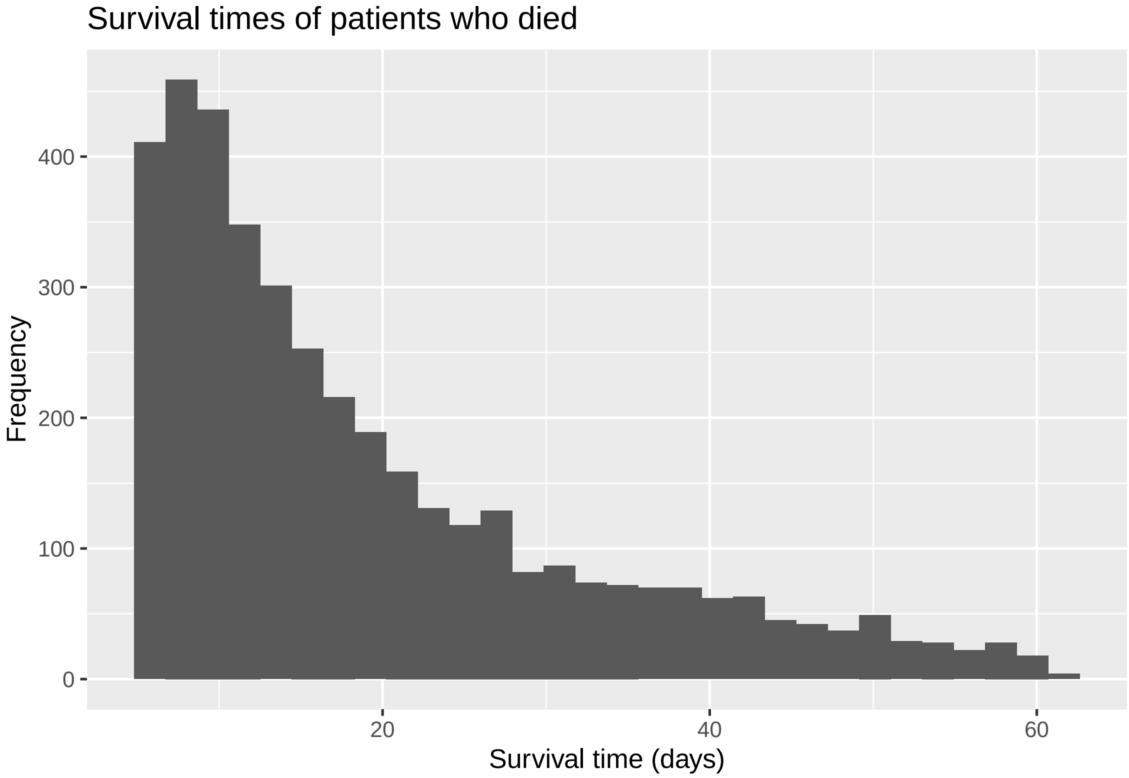

Figure 7.1: Survival histogram: Death in hospital (60-day mortality)

```
patient_dead <- patient[patient$PatientDied == 0, ]
ggplot(patient_dead, aes(x = event)) + geom_histogram() +
  xlab("Discharge time (days)") + ylab("Frequency") +
  ggtitle("Discharge waiting times of patients who survived")
```

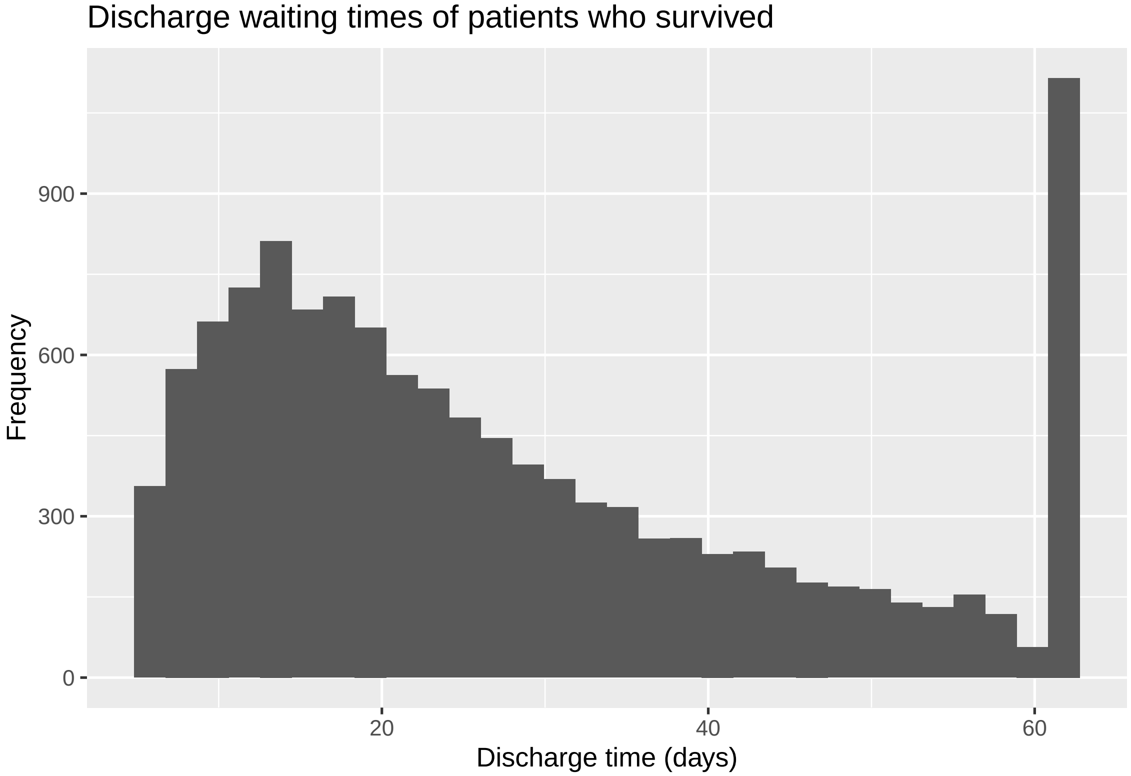

Figure 7.2: Survival histogram: Discharge from hospital (censored at 60 days)

For discharge, though, there is also mass at the end of the observed follow-up. These are essentially these patients with extremely late discharge events.

Significantly more men have been observed compared to women, which is not surprising from a clinical point of view. With a median age of 62 and a left-skewed distribution, we deal with a rather old population. The majority of cases are grouped in the admission category “medical” meaning simply they are not related to surgery. The diagnosis ID indicates that a large number of cases refer to respiratory problems and also cardiovascular illnesses are very much represented. The Apache II score is on average around 22 (median 22 and mean 22.08). We want to put this number into context. The predicted death rate can be derived using the ApacheII score as input to the logistic response (Knaus et al. (1991)):

$$\frac{\exp(-3.517 + ApacheII * 0.146)}{1 + \exp(-3.517 + ApacheII * 0.146)} \tag{7.1}$$

A score of 22 is associated with an anticipated mortatily probability of 42 percent. This means that the patients in this data set are – on average – critically ill. To be precise, the average predicted mortality in the data set is 36 percent, though, if we interpret the logistic regression accordingly. However, as we only investigate patients who need to be mechanically ventilated and need artificial nutrition this is no big surprise. Still, there is a gap of 11 percentage points in hospital mortality according to the event indicator.

```r
merged_day1 <- merged[merged$Study_Day == 1, ]
ggplot(merged_day1, aes(x = calCat, y = ApacheIIScore)) +
  geom_boxplot() + ylab("Apache II score") + xlab("Calorie category") +
  ggtitle("Distribution of the Apache II score w.r.t. caloric intake on admission day")
```

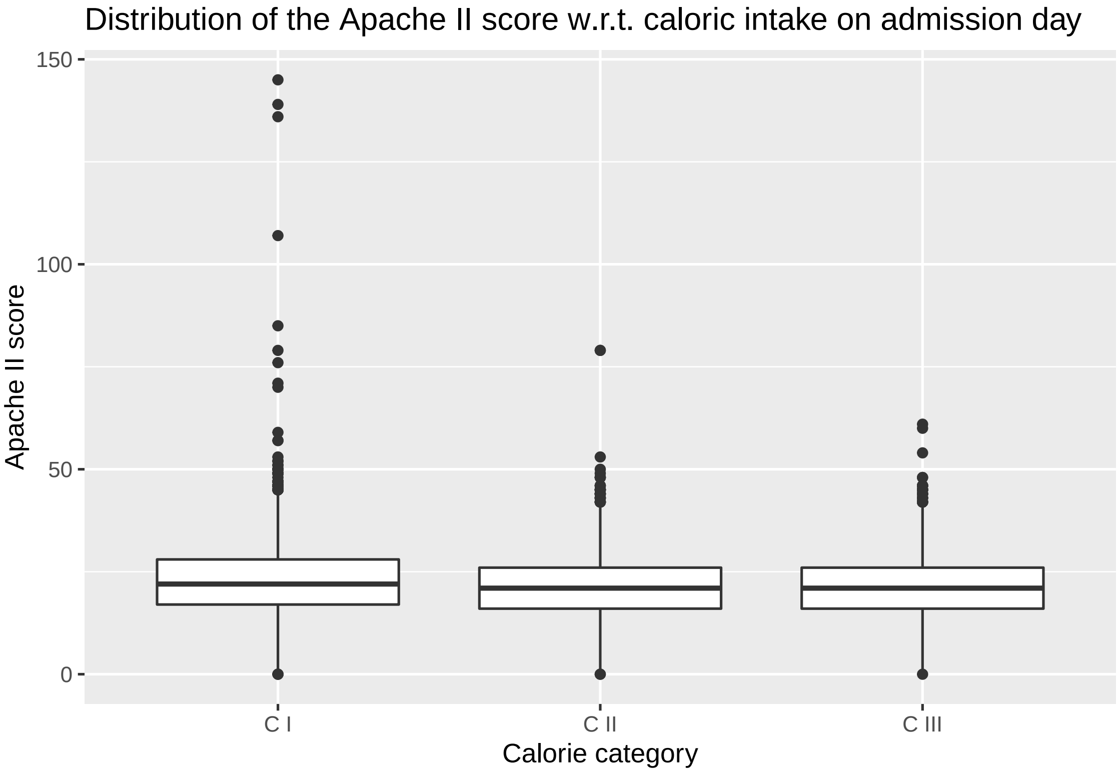

Figure 7.3: Box plots: Distribution of diet categories (calories) and the Apache II score

When investigating the caloric category on the day of admission, we see that all three categories are similarly distributed w.r.t. to the Apache II score. However, there seem to be many more of the most critically ill patients in the lowest category.

For protein intake, we observe the same pattern:

```r
ggplot(merged_day1, aes(x = proteinCat, y = ApacheIIScore)) +
  geom_boxplot() + ylab("Apache II score") + xlab("Protein category") +
  ggtitle("Distribution of the Apache II score w.r.t. protein intake on admission day")
```

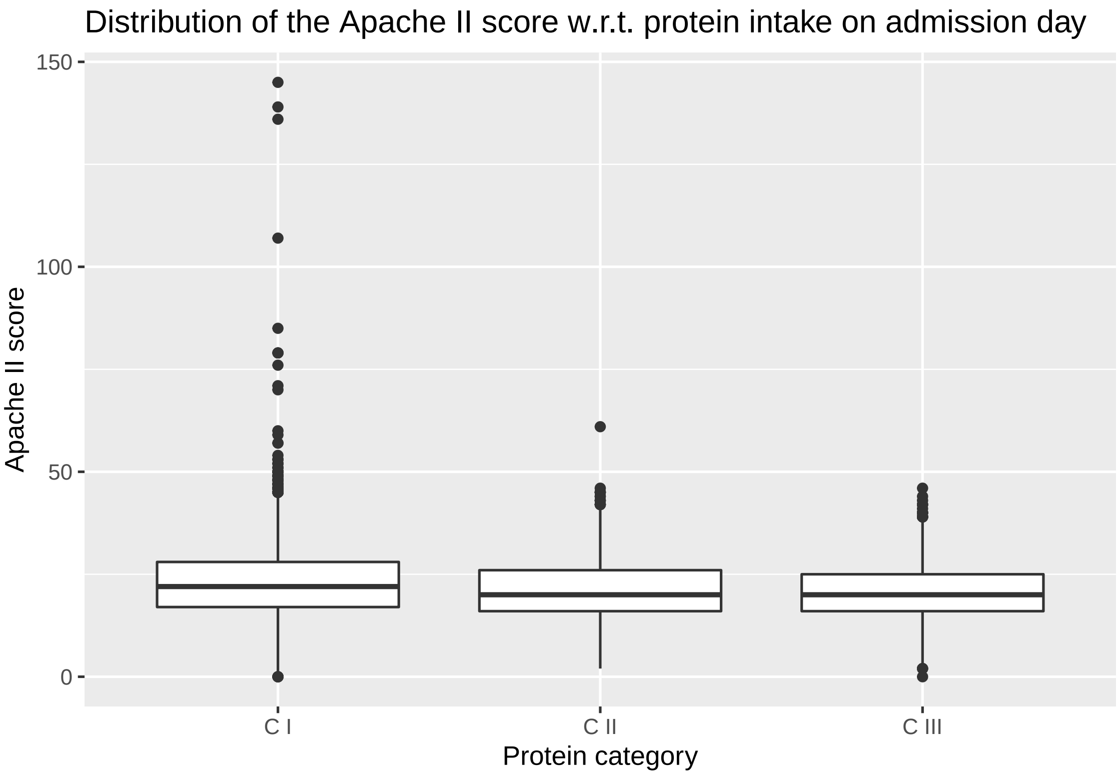

Figure 7.4: Box plots: Distribution of diet categories (protein) and the Apache II score

The covariates for the prescribed caloric and protein intake seem reasonably high and have low variance. These covariates are also highly correlated. The caloric intake and the protein intake (in percent of the prescribed intake) have a Pearson correlation of 0.83

Nutrition covariates are reported daily. We do not account for the special data structure (clustered by individuals) when reporting these patient days. For both, caloric and protein adequacy the summary statistics indicate that many patients did not receive any nutrition or – if any – very few. This can be reasoned by the fact that some critically-ill patients simply are too ill to be fed (Bender (2018)) or some ICU guidelines suggest no feeding for some patients (Heyland et al. (2011)). As propofol is supplied in a solution that contains fats, these critically-ill patients may have received propofol at some days as the only source of caloric intake.

For both, caloric and protein intake, we derive categorical features in the previously described manner. If a patient had an additional oral intake, she is assumed to move one caloric category up. This is because oral intake is only reported via an indicator. For protein intake, there is no moving up.

One can see in table 7.3 that there is a substantial difference between protein and caloric adequacy categories – despite a very strong dependence between the original raw covariates. If oral intake would also affect the protein intake, covariates in the same way as caloric intake, the distribution of patient days would look a bit more similar. The associated information is stored in `protCat`.

```
merged$protCat <- "C I"
merged$protCat <- ifelse(merged$protCat2 == 1, "C II", merged$protCat)
merged$protCat <- ifelse(merged$protCat3 == 1, "C III", merged$protCat)
```

```
merged$protCat <- as.factor(merged$protCat)
```

```
data.frame("Count" = summary(merged$protCat))
```

	Count
C I	70877
C II	50624
C III	55012

However, the definitions of protein and caloric adequacy are independent of each other and operate on a different scale. Both are reasoned from a medical point of view concerning nutrition adequacy. From the summaries one can infer that typically the nutrition received by patients contains under proportional amounts of protein.

7.3 Modeling

To investigate the effect of artificial nutrition on survival in, and discharge from the hospital, we employ the main specification of Bender et al. (2018b) with their exposure-lag-response associations. Refer to chapter three for an exhaustive discussion of these effects. We processed the data in the same fashion as Bender et al. (2018b) into three categories for both caloric and protein intake. We also assume the same dynamic lag-lead structure as Bender et al. (2018b).

```
library(pammtools)
ll_fun = function(t, tz) { t >= tz + 4 & t <= tz * 3 + 12  }
ll <- get_laglead(0:30, tz = 1:11, ll_fun = ll_fun)
gg_laglead(ll) + theme(text = element_text(size = 12),
                       axis.text.x = element_text(angle = 90, hjust = 1))
```

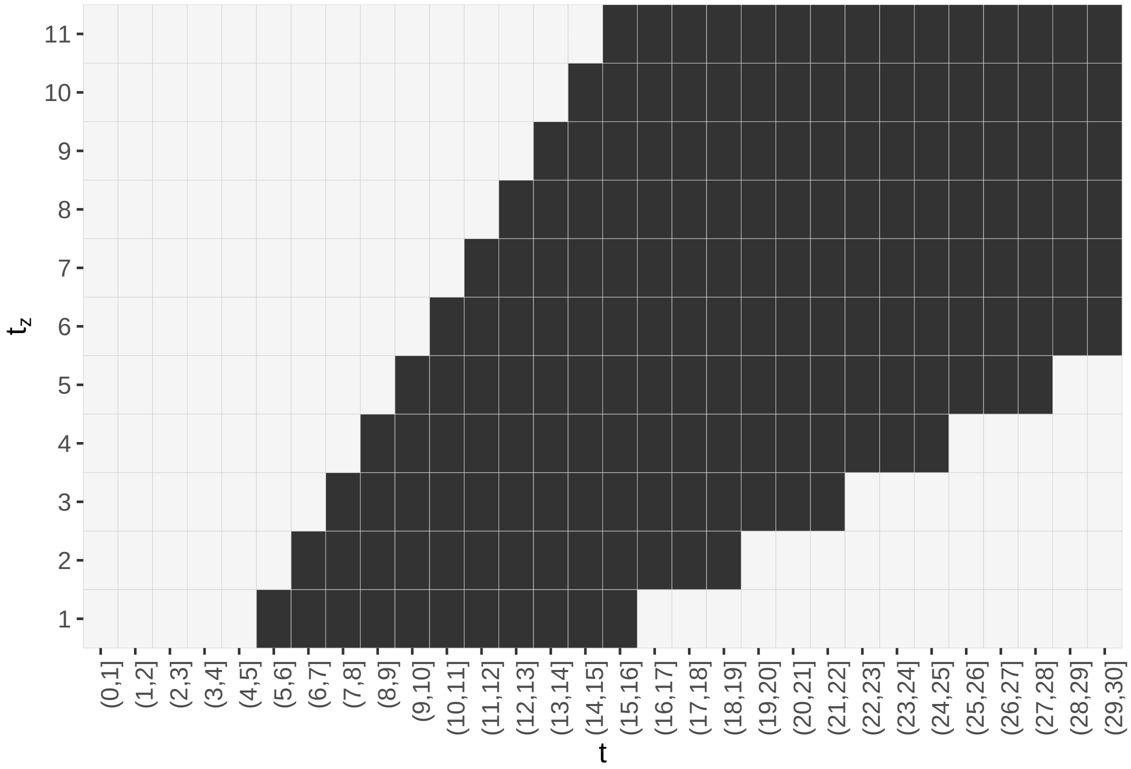

Figure 7.5: Lag-lead window specification in the analysis. The lag-lead window has a minimal lag of 4 days and is dynamically increasing.

There is a four-day long lag. Later study days increasingly affect more follow-up days. To be precise, each additional study day increases this window by three days.

In contrast to Bender et al. (2018b), we also analyse – next to the risk of dying in the hospital – the "risk" of being discharged as a competing event. Furthermore, we will also investigate protein intake. Still, the model notation of Bender et al. (2018b) can be borrowed:

$$\begin{aligned}
\log(\lambda_i(t|X)) = \\
&\lambda_0(t) + \beta_{year} * x_{i,year} + \beta_{diag} * x_{i,diag} + \\
&\beta_{admission} * x_{i,admission} + \beta_{gender} * x_{i,gender} + \beta_{MV} * x_{i,MV} + \\
&\beta_{propofol} * x_{i,propofol} + \beta_{oral} * x_{i,oral} + \beta_{parenteral} * x_{i,parenteral} \\
&\beta_{Apache} * x_{i,Apache} + \beta_{Apache:t} * (x_{i,Apache} * t) + \\
&f_{age}(x_{i,age})t + \\
&f_{BMI}(x_{i,BMI})t + \\
&\sum_{t_e} g_{c_{II}}(t_e, t) + \sum_{t_e} g_{c_{III}}(t_e, t) + \\
&b_{l_i}
\end{aligned} \tag{7.2}$$

where:

- $\lambda_i(t|X)$ is the log hazard of failing. $\lambda_i(t|X)$ either refers to the risk of dying within a 30-day follow-up in the hospital or being discharged in that period.
- $\lambda_0(t)$ is the (log) baseline hazard for the respective hazard just described.
- $\beta_{year} * x_{i,year}$, $\beta_{diag} * x_{i,diag}$, $\beta_{admission} * x_{i,admission}$ and $\beta_{sex} * x_{i,sex}$ are a fixed effect for each study year, the diagnosis and admission category and the patients' sex.
- $\beta_{MV} * x_{i,MV}$, $\beta_{propofol} * x_{i,propofol}$ $\beta_{oral} * x_{i,oral}$ and $\beta_{parenteral} * x_{i,parenteral}$ are linear effects for the number of days with mechanical ventilation, propofol administration, oral intake and parenteral intake between the second and the fourth study day (i.e. three days).
- $\beta_{Apache} * x_{i,Apache} + \beta_{Apache:t} * (x_{i,Apache} * t)$ is a linear, linearly time-varying effect for the ApacheII score.
- $f_{age}(x_{i,age})t$ and $f_{BMI}(x_{i,BMI})t$ are smooth, smoothly time-varying effects for the patients' age and their BMI.
- $g_{C_{II}}(t_e, t)$ and $g_{C_{III}}(t_e, t)$ are the cumulative effects for the both upper (C II and C III) nutrition categories. For a detailed derivation, refer to chapter 3. The effects represent – depending on the model which is fit – either caloric intake or protein intake.
- b_{l_i} is a random effect for each single ICU.

7.3.1 Model set ups

We study 8 different models in total: For both, caloric intake and protein intake (2), we investigate the subdistribution and a cause-specific hazards model (2 x 2). For each competing risks model, we analyse both competing risks (2 x 2 x 2 = 8).

7.3.1.1 Model 1 - Caloric intake

The caloric intake model features four different models. The first model presented is the same as in Bender et al. (2018b). (However, we use more data.) The remaining 3 models are new and provide the investigation of competing risks lacking in Bender et al. (2018b).

7.3.1.1.1 Subdistribution hazards

The first specification is a subdistribution hazards model. We achieve a subdistribution hazards model by suggesting proper censoring times. As there is only administrative censoring in the data, the proposed model is derived by setting survival times to the moment of administrative censoring. Since the data features time-dependent covariates, missing time-dependent values **may** be an issue. However, we accounted for these missing values already by carrying-on missing information in the nutritional intake with a C III day. Thus, there are no missing time-dependent values. Nonetheless, this (strong) assumption is examined later.

The `ped` data set can be defined via `pammtools` with `as_ped_cr_sh()`. However, we decided to model each sub-model of the design of a competing risk separately as our models are computationally expensive. Thus, we only make use of `pammtools::as_ped()`. Additionally, we add the cumulative effects not via `cumulative()` in `pammtools:as_ped()`. This is because the cumulative data preprocessing in this scenario needs special treatment (the effect is trivariate) which is currently not optimally supported in `pammtools::cumulative()`. Thus, we make use of the custom-made `add_cumulative_eff_vec()` function.

```
patient_1_A <- patient
patient_1_A$event[patient_1_A$PatientDischarged == 1] <- 61L
ped_1_A <- as_ped(
  data    = patient_1_A,
  formula = Surv(event, PatientDied) ~ Year + DiagID2 + AdmCatID + Gender +
    ApacheIIScore + BMI + Propofol2_4 + inMV2_4 + OralIntake2_4 + PN2_4 + Age +
    CombinedicuID + icuByDummy,
  cut     = 0:30, id = "CombinedID")
ped_1_A <- ped_1_A %>%
  add_cumulative_eff_vec(daily, "calCat2", "calCat3", LL = ll)
ped_1_A$int_mid <- 0.5 * (ped_1_A$tstart + ped_1_A$tend)
ped_1_A <- ped_1_A[ped_1_A$tend > 5, ]
```

`patient_1_A$event[patient_1_A$PatientDischarged == 1] <- 61L` performs administrative censoring in the sense of Fine and Gray (1999). `ped_1_A$int_mid <- 0.5 * (ped_1_A$tstart + ped_1_A$tend)` adds interval mid points for the effects. `ped_1_A <- ped_1_A[ped_1_A$tend > 5, ]` deletes the first 5 days of the follow-up. This only speeds up computation as they are irrelevant to the model.

The fitting of the model is comparatively trivial as a normal Poisson regression.

```
m_1_A <- bam(formula = formula, data = ped_1_A,
             family = poisson(), offset = offset)
```

Note that the actual model is loaded from the repo of this gitbook. A real-time computation as the book compiles would be too expensive.

For the evaluation of the model, we make use of the same graphical analysis as Bender et al. (2018b). They compare the following diets with each other in six different pairwise comparisons.

Table 7.4: Comparison of diets.

Comparison	Z_1	Z_2
A	Days 1-11: C I	Days 1-4: C I, Days 5-11: C II
B	Days 1-11: C I	Days 1-11: C II
C	Days 1-4: C I, Days 5-11: C II	Days 1-11: C II
D	Days 1-11: C I	Days 1-11: C III
E	Days 1-11: C II	Days 1-4: C II, Days 5-11: C III
F	Days 1-11: C II	Days 1-11: C III

For each comparison, they compute the differences in hazard rates. Both, Z_1 and Z_2 are *hypothetical* diets. The associated hazard originates from the prediction of the model. For better visualisation the difference is transformed via the exponential function so that differences can be interpreted as factor differences.

$$e_j = \frac{\lambda\left(\tilde{t}_j|\mathbf{z}_2\right)}{\lambda\left(\tilde{t}_j|\mathbf{z}_1\right)} = \exp\left(\mathbf{g}_{z_2} - \mathbf{g}_{z_1}\right) \tag{7.3}$$

The resulting ratio is centered around 1. Z_1 is always the diet with less nutrition compared to Z_2. Thus, a ratio below 1 means that more nutrition reduces the risk of the event. A ratio larger than 1 is associated with an increased hazard. We display the ratio (solid line) with its 95% confidence bands (dashed lines).

```
comp_frames <- make_six_frames(m_1_A, patient, daily, ll_fun,
                               var1 = "calCat2", var2 = "calCat3",
                               type = "1", effect = "calCat")
six_plots(comp_frames[[1]], comp_frames[[2]], comp_frames[[3]],
          comp_frames[[4]], comp_frames[[5]], comp_frames[[6]])
```

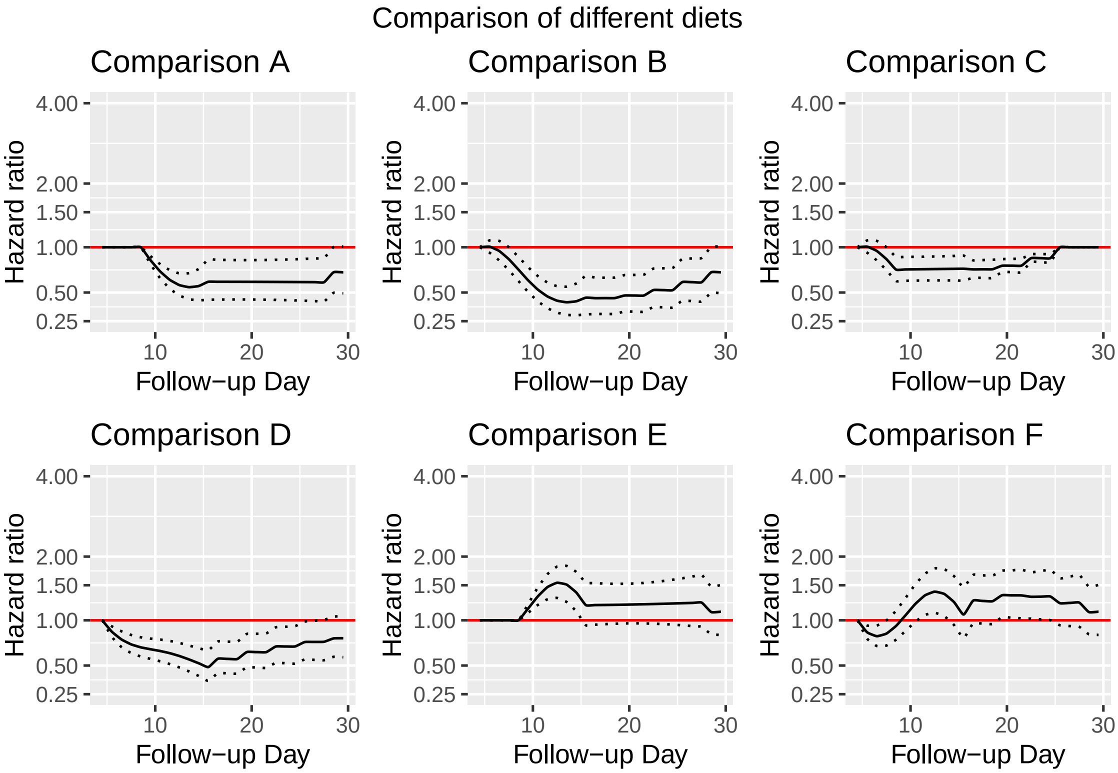

Figure 7.6: Estimated caloric effects on hospital mortality - Subdistribution hazards model

In figure 7.6 we study the estimated effects of caloric intake on survival in hospital. We evaluate the previously estimated subdistribution hazards model. We observe that for pairwise comparisons A, B, C, and D there is a significantly negative effect on the hazard ratio. For E and F the effect is significantly positive for most days of the follow-up. Comparisons A, B, C, and D involved at least to some extent a diet based on C I nutrition. Moving away from this diet seems to be beneficial in all comparisons listed here. However, comparisons E and F involve moving from a (strictly) C II diet to a (partly/strictly) C III diet. This does not seem to reduce the risk of dying in hospital but rather increases it. However, comparison D still shows that moving from a hypocaloric diet (strictly C I) to a high-caloric diet (strictly C III) is beneficial.

our results are almost identical to Bender et al. (2018b). All effects have the same shape. However, most effects appear somewhat stronger. Furthermore, we find that the confidence bands decreased in our approach for all comparisons. Thus, the effects described by Bender et al. (2018b) are reinforced in our data with higher certainty. Additionally, the effects of Comparison E and F are now significant compared to Bender et al. (2018b) who find them to be insignificant.

While we are mainly interested in the event “death in hospital”, the competing event “discharge” is analysed as well to study the competitiveness of the events.

```
patient_1_B <- patient
patient_1_B$event[patient_1_B$PatientDied == 1] <- 61L
ped_1_B <- as_ped(
  data     = patient_1_B,
  formula = Surv(event, PatientDischarged) ~ Year + DiagID2 + AdmCatID +
    Gender + ApacheIIScore + BMI + Propofol2_4 + inMV2_4 + OralIntake2_4 +
    PN2_4 + Age + CombinedicuID + icuByDummy,
  cut      = 0:30, id = "CombinedID")
ped_1_B <- ped_1_B %>% add_cumulative_eff_vec(daily, "calCat2", "calCat3",
                                              LL = ll)
ped_1_B$int_mid <- 0.5 * (ped_1_B$tstart + ped_1_B$tend)
ped_1_B <- ped_1_B[ped_1_B$tend > 5, ]

m_1_B <- bam(formula = formula, data = ped_1_B,
             family = poisson(), offset = offset)
```

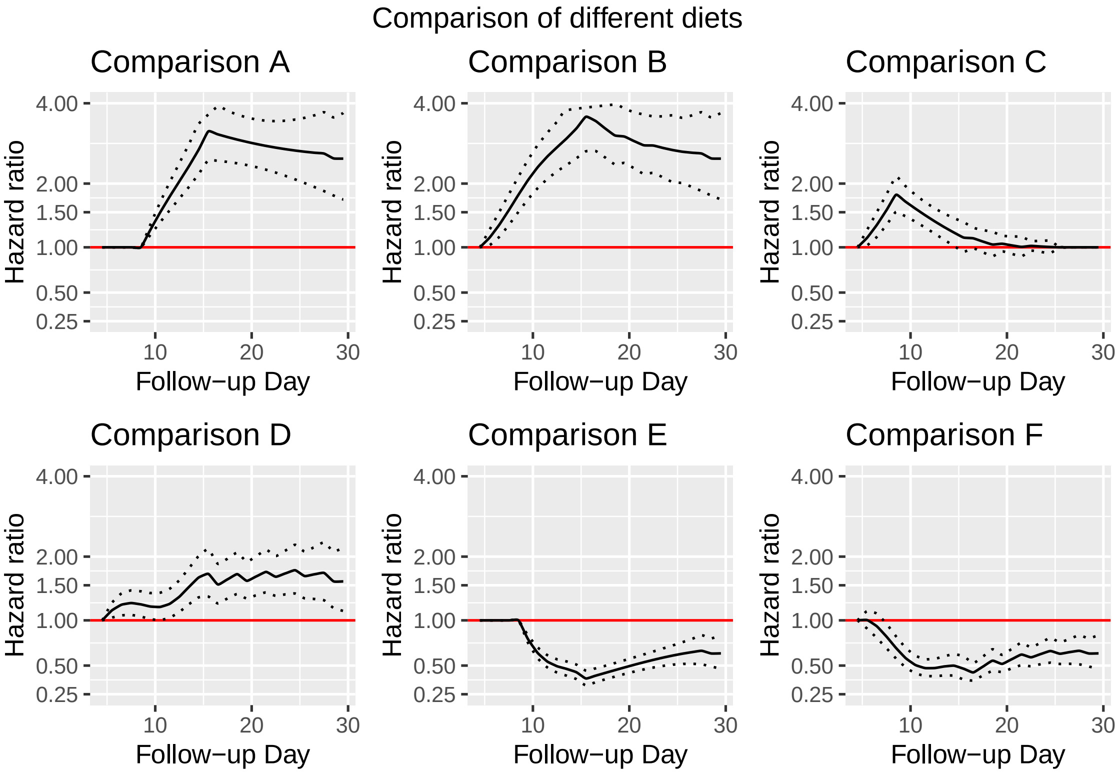

Figure 7.7: Estimated caloric effects on discharge from hospital - Subdistribution hazards model

We observe that all effects are the opposite effects of the previous model where death has been modeled. Loosely speaking, the discharge process is accelerated if one moves from hypocaloric nutrition away. The discharge process is slowed down when moving even further to high caloric food.

7.3.1.1.2 Cause-specific hazards

Before evaluating the subdistribution model, we want to present the results of the cause-specific hazards model. Both models are evaluated vis-à-vis later on. We perform the same post-hoc analysis as for model A.

The cause-specific model is created very similarly. One only does *not* alter the discharge and survival time. The relevant code can be accessed within the repository of this thesis.

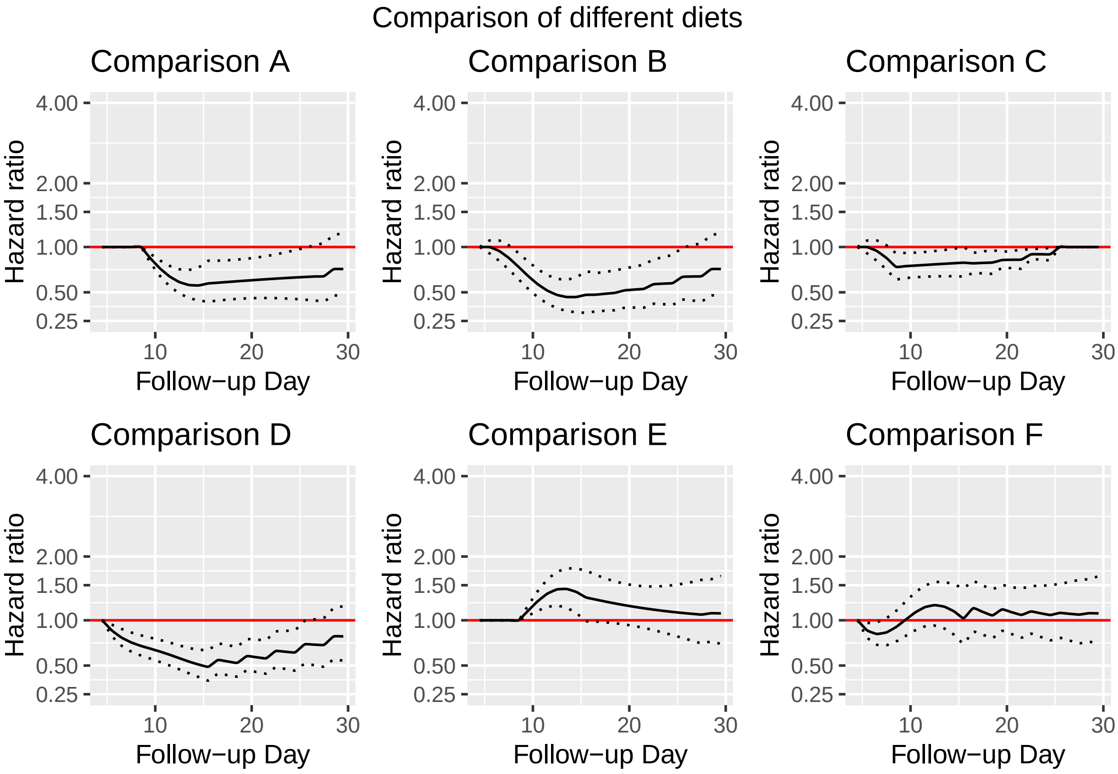

Figure 7.8: Estimated caloric effects on hospital mortality - Cause-specific hazards model

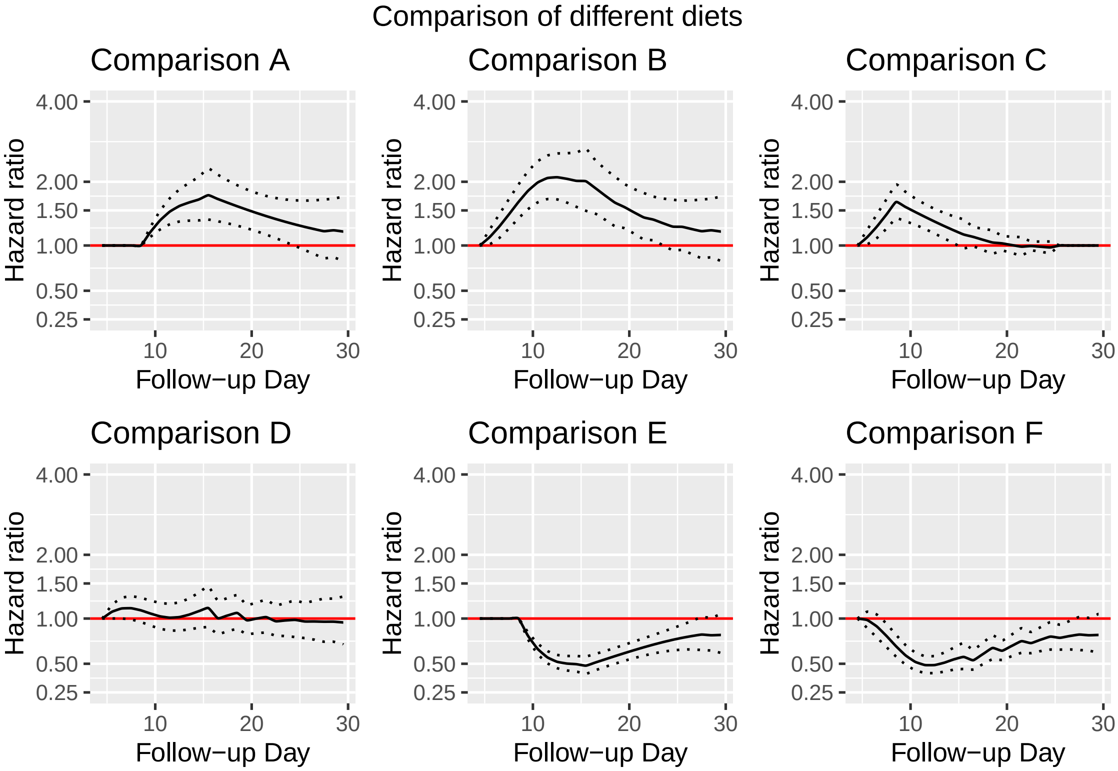

Figure 7.9: Estimated caloric effects on discharge from hospital - Cause-specific hazards model

The effects in figure 7.8 and 7.9 look similar compared to their subdistribution counterparts in figures 7.6 and 7.7. Some effects are less strong, though. This is especially true for the modeling of the discharge event. Most comparisons are qualitatively identical to the subdistribution ones.

The model suggests that high and medium-caloric diets dominate hypocaloric diets in terms of extended survival times. The model suggests that high-caloric diets are associated with decreased survival times compared to medium-caloric diets. The difference between these two diets tends to be insignificant, though.

The estimated effects indicate that decreased discharge times are associated with both high and hypocaloric nutrition compared to medium-caloric diets. Concerning the discharge both, high and hypocaloric diets seem to have similarly disadvantageous effects.

7.3.1.1.3 Evaluation

Vis-à-vis, the analysis of cause-specific and subdistribution hazards suggests very similar estimated effects for most comparisons.

The cause-specific model tends, though, to measure smaller and less significant effects. It seems that estimating the CIF and cumulative hazards point to very similar inferences. This argues for a weak dependence between the event types (discharge vs. death), and hence low competitiveness. The underlying mechanisms according to which caloric intake affects the different event probabilities work separately from one another.

Both models suggest that moving from hypocaloric diets away to medium-caloric diets increases survival time and decreases discharge time. The two models disagree on the effect of moving even further to a high-calorie diet. The subdistribution hazards model suggests that survival times are shortened and discharge times extended. However, the cause-specific model assesses this differently: higher-caloric diets do not seem to be significantly associated with shorter survival times. In the cause-specific model, the only significant comparison involving a high-caloric diet remains comparison D (strictly C I vs. strictly C III). This comparison, however, still points to an extended survival time.

For discharge, both models also have similar effect shapes. The subdistribution hazards model, however, tends to feature stronger effects. The only comparison which the two models disagree on is comparison D. The subdistribution hazards model finds that a strict C III diet compared to a strict C I diet accelerates hospital discharge. The cause-specific model does not support this idea, though.

Typically, we are trying to identify causal effects in this analysis. As the literature suggests (see chapter 4), this proposes the use of the cause-specific hazards model. However, the subdistribution hazards model assists in confirming the cause-specific findings throughout. Both models essentially tell the same story. Thus, we can be very agnostic at this point. From the cause-specific estimates, we draw the same conclusions as Bender et al. (2018b): Survival time increases when moving away from hypocaloric diets to medium or higher-caloric diets. However, based on the data, it does not pay off to increase the caloric intake, even more, coming from a medium-caloric diet.

Furthermore, we find that moving from a hypocaloric diet to a medium-caloric diet decreases discharge time. Hypocaloric diets and high caloric diets seem to have very similar effects so that moving from medium-caloric diets to higher-caloric diets increases discharge time again.

The inference from this analysis is that a medium-caloric diet is advantageous compared to different diets.

7.3.1.2 Model 2 - Protein intake

For protein intake, we use the very same procedure as for caloric intake. We refrain in this section entirely from displaying the associated code.

7.3.1.2.1 Subdistribution hazards

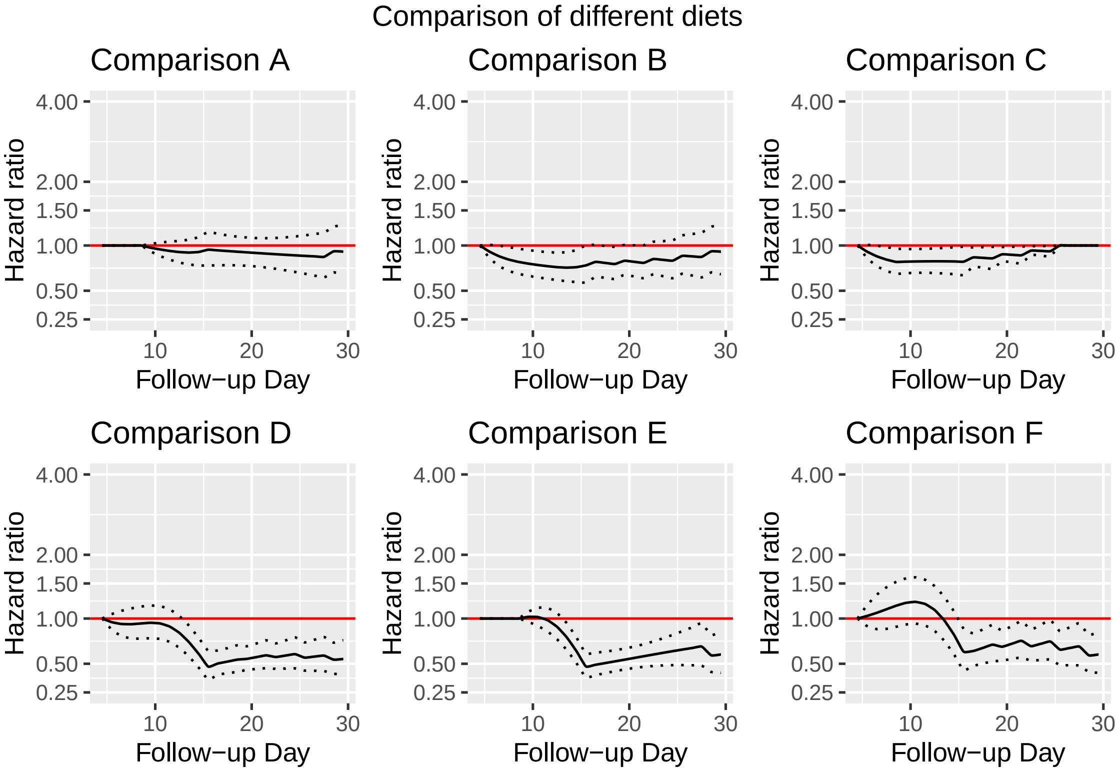

Figure 7.10: Estimated protein effects on hospital mortality - Subdistribution hazards model

We start with the effect of protein intake on survival times (figure 7.10). The subdistribution hazards model suggests that all effects are rather similar between the different diets in the short-run. Medium-caloric diets seem to have some (little but still significant) advantage in the short-run. However, the only strong effect is associated with moving to high-protein nutrition. While there is some evidence for medium-protein nutrition in the short-run, the results promote high-protein nutrition in the long-run. While these comparisons are not at all exhaustive, they indicate that an early intake of medium levels of protein which increases to higher levels after some days is associated with beneficial outcomes. Diets that administer high levels of protein on later study days are associated with significantly lower hazard rates.

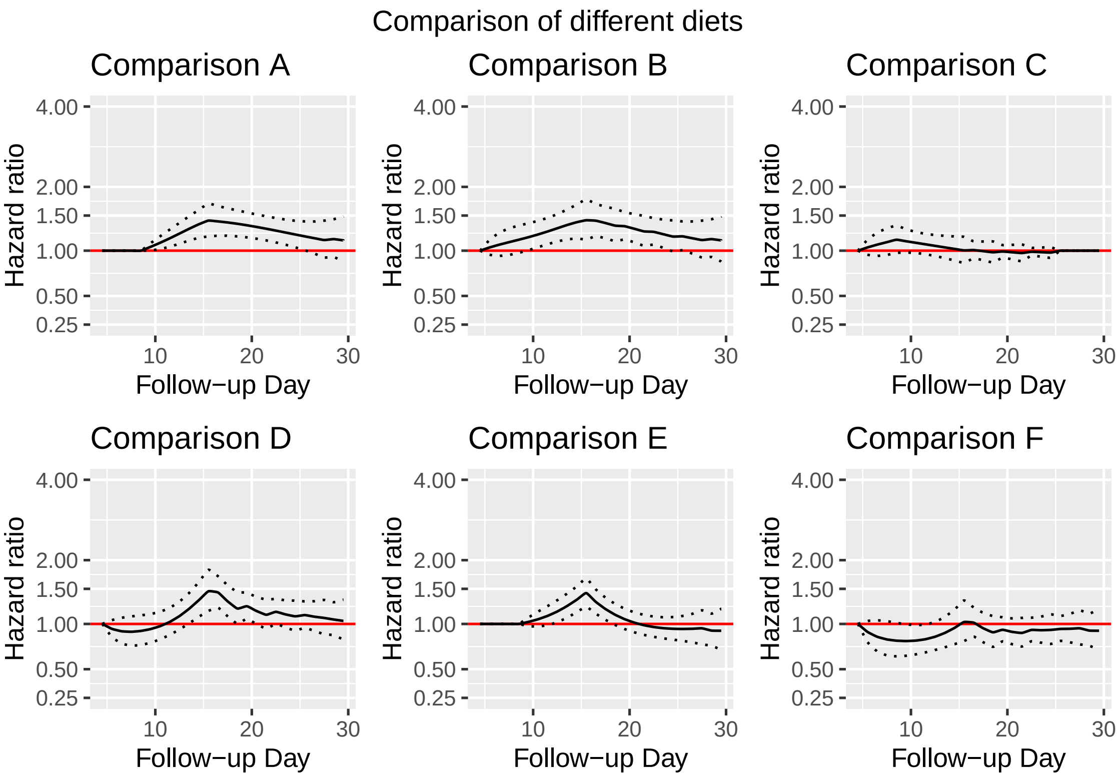

Figure 7.11: Estimated protein effects on discharge from hospital - Subdistribution hazards model

For discharge time (figure 7.11), the significant effects are mainly limited to medium-protein diets. Discharge time is shorter if medium-protein nutrition has been administered compared to all different diets. An involvement of higher-protein nutrition leaves ambiguous results: The effects do not clearly point to a significant difference between medium-protein and higher-protein discharge times.

7.3.1.2.2 Cause-specific hazards

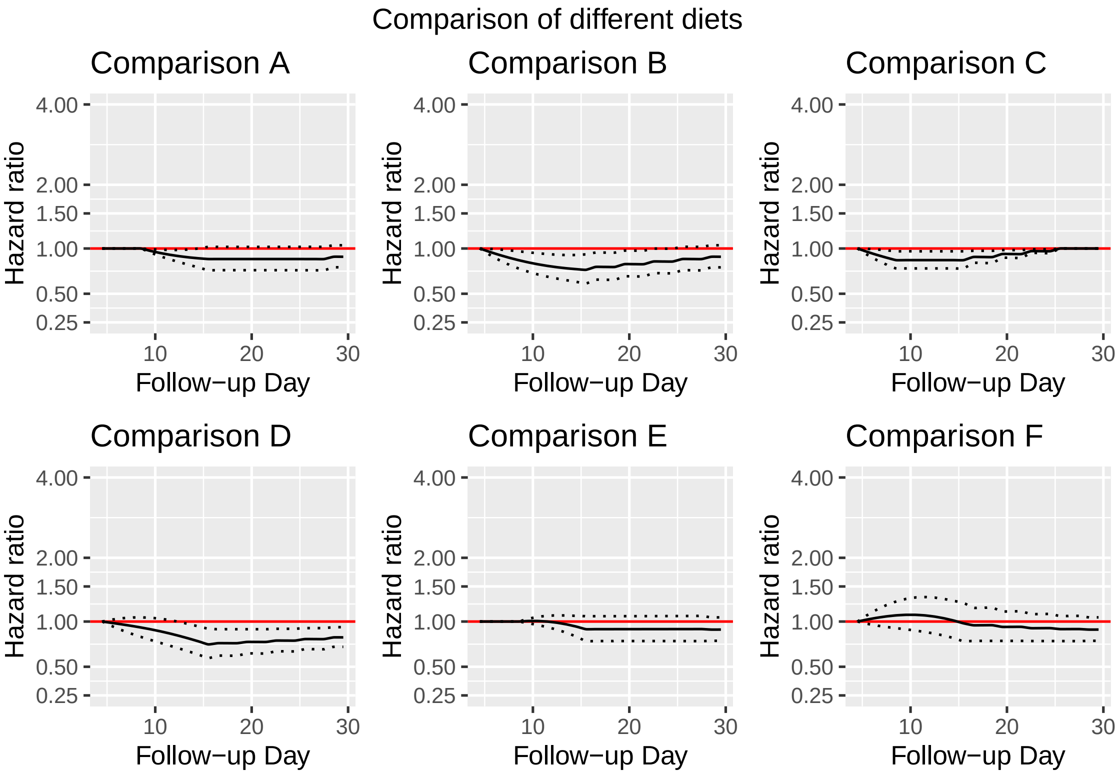

Figure 7.12: Estimated protein effects on hospital mortality - Cause-specific hazards model

Again the first model presented (in figure 7.12) is the one for survival times. The estimated effect of the cause-specific hazards model is less strong and implies different inferences than the subdistribution hazards model. While low-protein diets seem to be dominated by both, medium and higher-protein nutrition, the data finds the effects of medium and higher-protein nutrition to be very similar. Their differences are not significant.

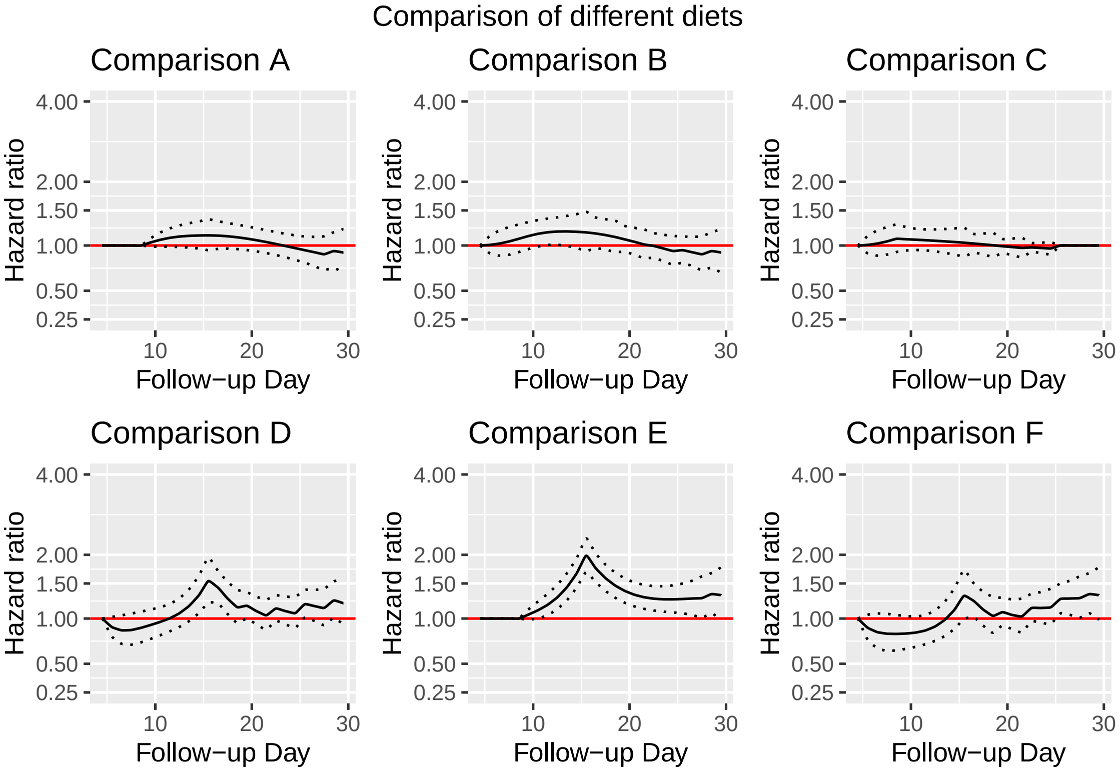

Figure 7.13: Estimated protein effects on discharge from hospital - Cause-specific hazards model

For discharge time (figure 7.13), most comparisons suggest that the differences between all categories of nutrition are not significant. However, a slight short-run effect of medium-protein nutrition and a long-term effect of high-protein nutrition can be observed. Thus, the only notable comparison is E, where this diet has been compared against a strict C II diet.

7.3.1.2.3 Evaluation

The analysis of the effect of protein intake is more ambiguous than for caloric intake. While the subdistribution hazards model especially promotes high-protein nutrition to extend survival times, the cause-specific model does not find any difference between medium and high-protein nutrition. Both models only agree within their tendencies of "more is better".

With respect to discharge time the models only agree on the positive short-term effect of medium-protein nutrition.

Synoptically, the approaches tend to see advantages of medium or higher-protein nutrition. This is true for both, extending the survival time and decreasing the discharge time. They disagree, though, which of these is eventually *better*.

Unlike for caloric intake, the cause-specific and subdistribution hazards model imply different effects. This has two potential reasons. In total, the estimated effects of protein tend to be smaller than for caloric intake. Hence they are harder to identify. Second, the mechanism of how the protein intake affects the two event types are not as independent as for caloric intake. At the same time, one needs to consider that both covariates have been constructed very differently as outlined previously.

For caloric intake, it was sufficient to be agnostic on the models: Both models essentially draw the same conclusions. In this case, we need to argue which model we think better fits the data situation. Bender et al. (2018b) assume that patients discharged from hospital survived at least until the end of the follow-up period. Their argument makes sense. Typically, these patients who recovered from their condition are released from the ICU and then from the hospital. However, this does not necessarily mean that patients released from the ICU/hospital recovered and will not die anymore due to the condition. At this point, an important fact has to be acknowledged. We can **only** model the event **death in hospital**. (And we did so in the whole analysis.) If one is precise, the individual discharged from the hospital is still under risk for dying (outside the hospital or after re-administration). For example, 1 in 5 UK ICU patients die within one year after discharge (Szakmany et al. (2019)).

If we limit on what we can actually model – this is the event **death in hospital** (or less importantly discharge from hospital) – the cause-specific hazards model will result in the appropriate estimates. This is because the marginal probability of dying in hospital (or being discharged) **only** depends on the actual current risk set. The approach of Bender et al. (2018b) leads to a more generic approach aiming to assess the event **death** in general.

Taking this into account, we can summarise the findings of the data analysis by the following: There is evidence that medium-protein nutrition extends survival times compared to low-protein nutrition. There are indications that survival times may be extended even further by high-protein nutrition. However, these effects are not significant. Discharge time is optimally reduced by medium-levels of protein on early study days and high-protein nutrition on later study days.

7.3.1.3 Robustness to different missing value methods

For individuals discharged from the ICU, their nutrition is not tracked anymore. For these patients Bender et al. (2018b) assume that a C III nutrition intake has been administered. They argue for this because patients discharged typically are healthier than the remaining ones. Thus, they would be fed normally in this condition. We also employ this assumption of Bender et al. (2018b) for both, caloric and protein intake.

In this section, we analyse how robust the model is w.r.t. this assumption.

We limit ourselves to the subdistribution hazards model for the evaluation of different carry-on methods. This is because this model is more affected by the missing value method. To be precise, we only investigate the subdistribution hazards model for dying in the hospital. We will focus on the caloric covariates. However, we also show how the protein effects are affected.

First, we review the carrying-on of the last observed value.

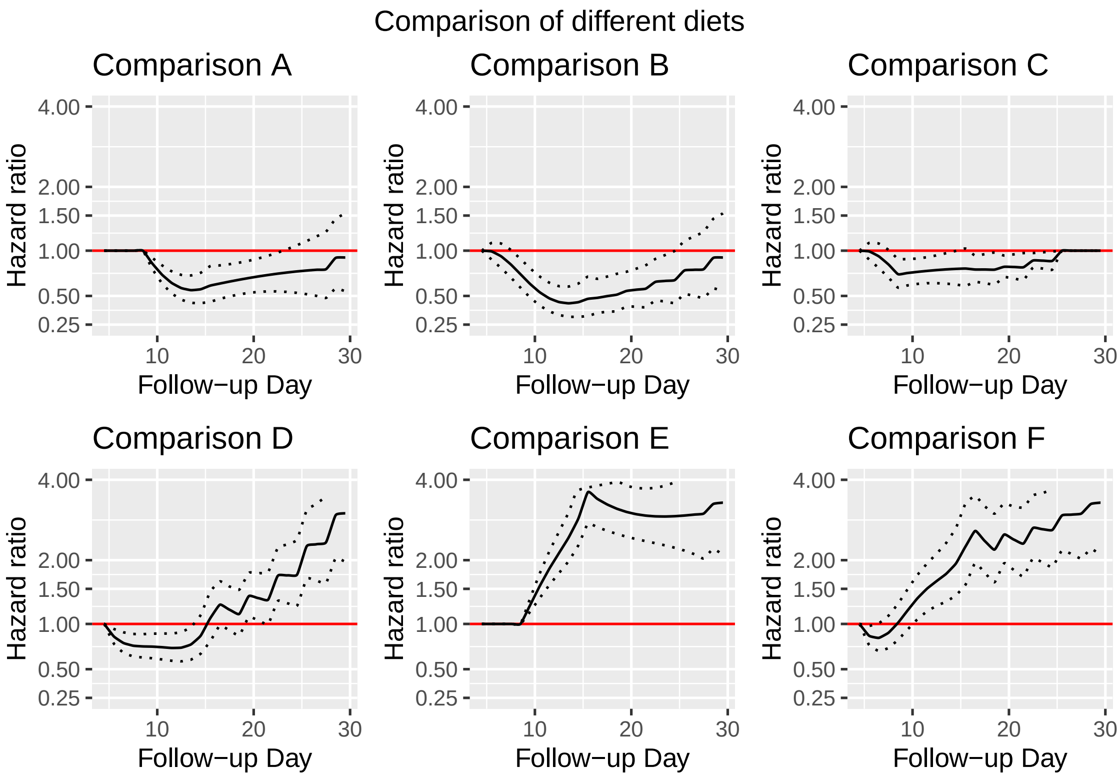

Figure 7.14: Estimated caloric effects on hospital mortality (Subdistribution hazards model) - Sensitivity analysis A

While the relative of C II seems to be stable, the effect of C III is very distorted. All comparisons involving a C III diet indicate that this sort of diet goes hand in hand with an extremely increased hazard of dying. One can observe that the estimated effects start diverging for later study days only. This is reasonable because typically later study days need to be imputed more often.

The next carry-on method is to use C I caloric intake for all missing values.

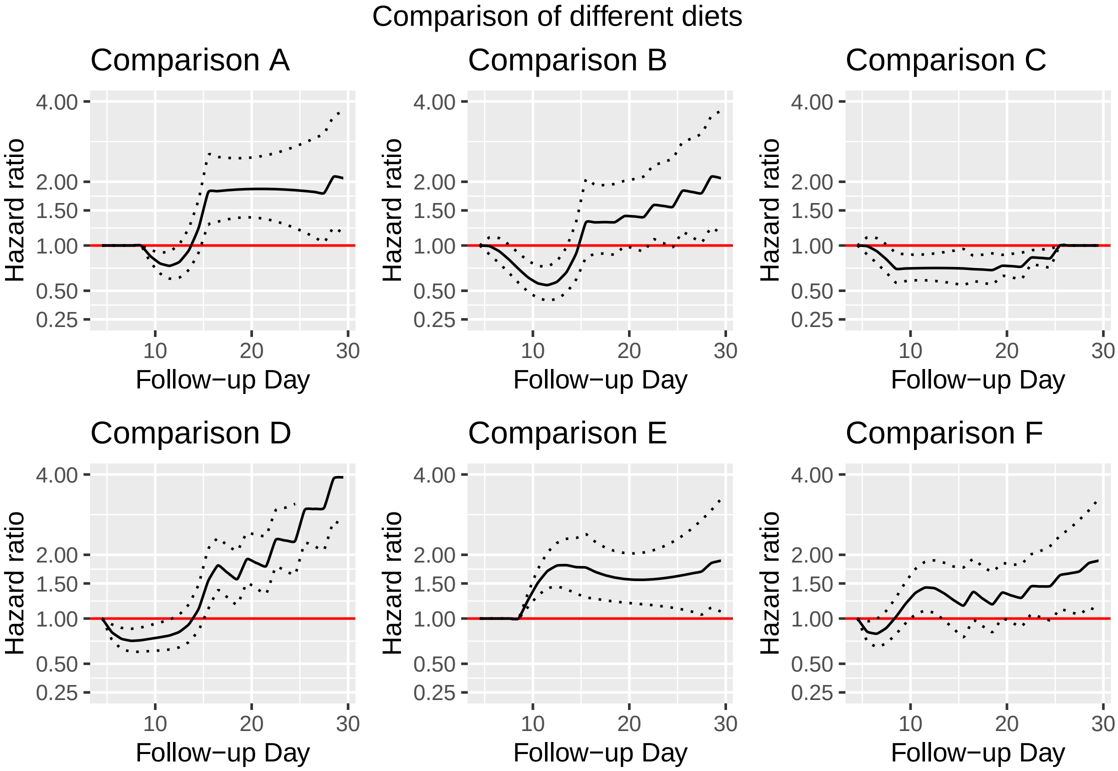

Figure 7.15: Estimated caloric effects on hospital mortality (Subdistribution hazards model) - Sensitivity analysis B

In this scenario, one can see that for later study days the relative effects between C II and C III are not very much affected. However, the effect of C I nutrition seems to strongly decrease for the later days of the follow-up.

We do the same for C II nutrition.

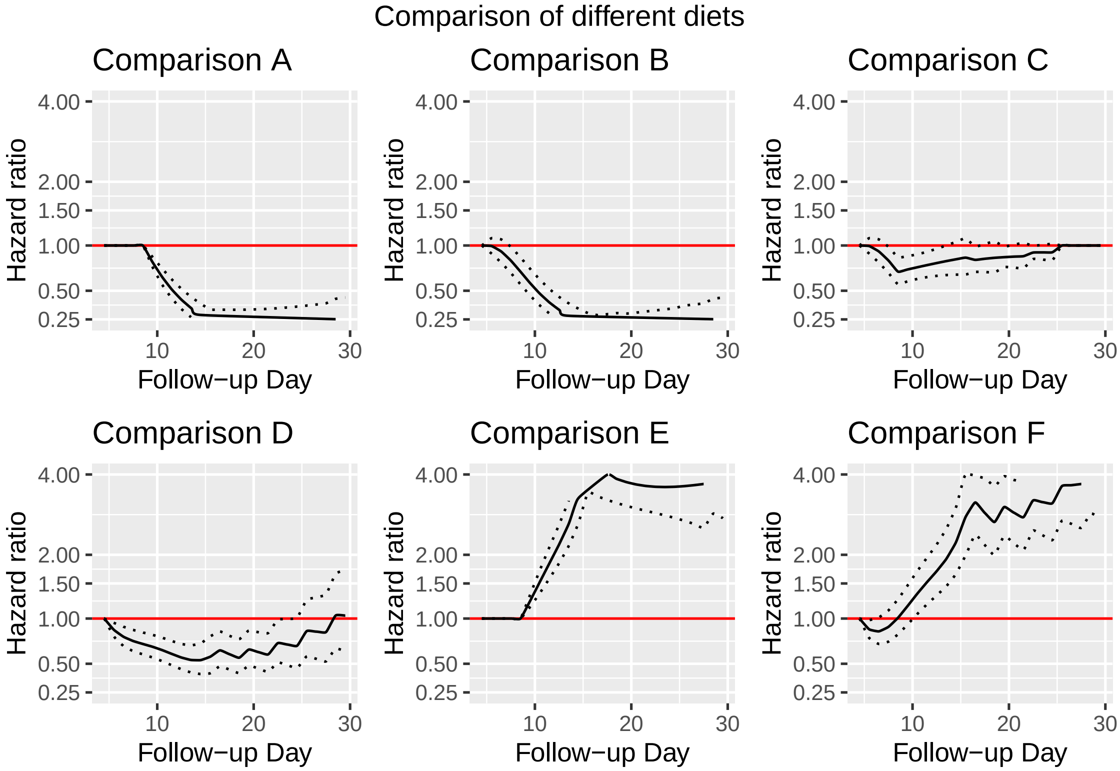

Figure 7.16: Estimated caloric effects on hospital mortality (Subdistribution hazards model) - Sensitivity analysis C

Here we can study the very same effect as before: the effect of the category to which missing protocols have been set becomes extremely negative.

The standard specification was reasonable and very well-argued by Bender et al. (2018b). The first sensitivity analysis (figure 7.14) has been the most realistic one next to the standard specification. The others are very extreme. However, they visualise the nature of these missing protocols very well. A missing protocol indicates a lower mortality risk. Additionally, they show how reliant the model is on the assumption of Bender et al. (2018b).

In our data we see that only 10 percent of these with incomplete protocols (and having survived at least 12 days) actually died in hospital – compared to over 23 % considering all patients. Of course, this is to some extent tautological: Missing protocols are partly induced by the fact that patients are discharged (i.e. they survived).

Thus, we analyse what happens if these patients with incomplete protocols are dropped. These are 6426 patients and hence a significant portion of the original data.

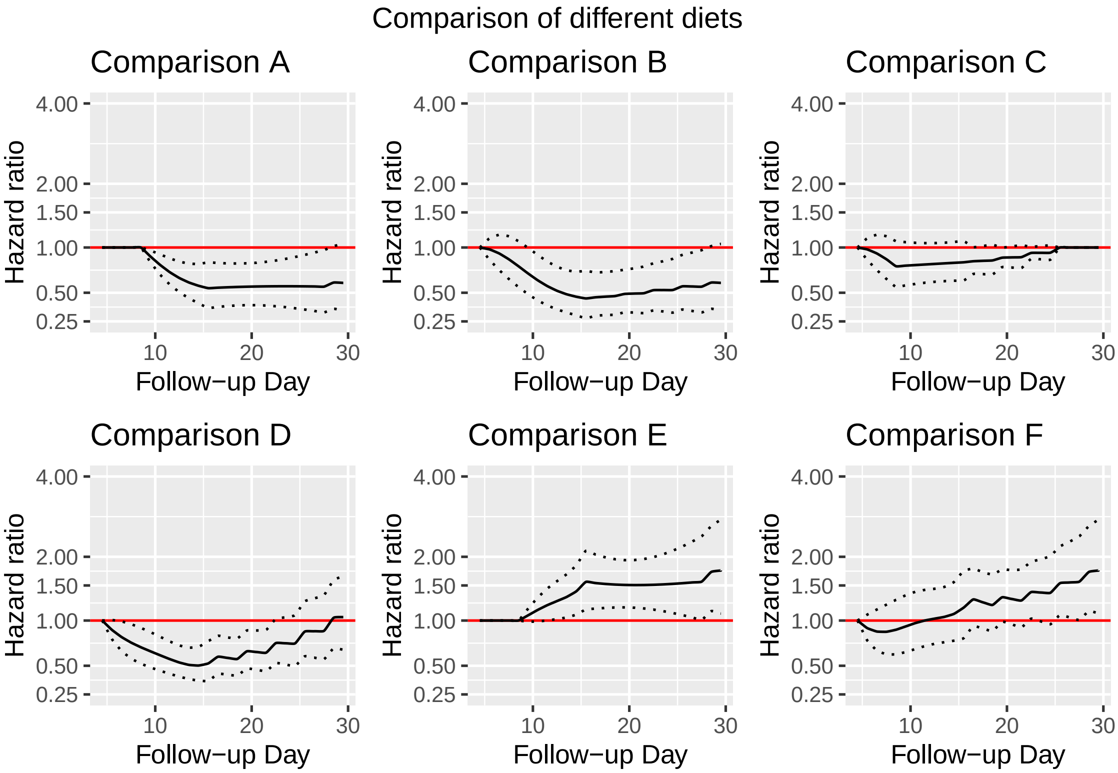

Figure 7.17: Estimated caloric effects on hospital mortality (Subdistribution hazards model) - Sensitivity analysis D1

Doing so, one receives a similar picture as for the main specification. However, now – due to significantly less data – the confidence bands are larger for all effects. Even when ignoring patients with incomplete protocols, the qualitative findings would be very similar to the ones of the main specification.

We also analyse protein intake in the same fashion.

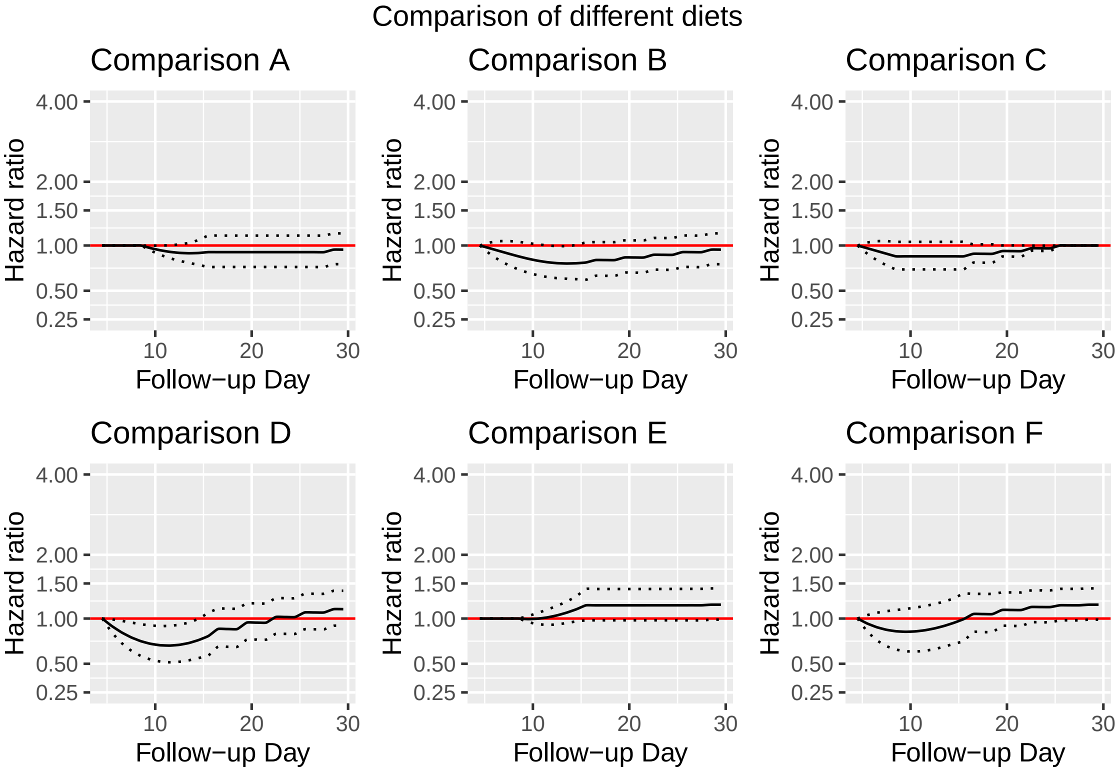

Figure 7.18: Estimated protein effects on hospital mortality (Subdistribution hazards model) - Sensitivity analysis D2

Here, one can draw the same picture: Effect estimations are similar. But the effect for C III seems to be smaller.

Sensitivity analysis A (figure 7.14) outlines that especially the effect category III is affected by altering the imputation assumption. Sensitivity analysis B and C (figures 7.14 and 7.15) state that this is since typically healthier individuals have imputed protocols. Sensitivity analysis D1 shows that neglecting all patients with missing protocols (typically very healthy individuals) has only little effect on the estimated effect for C III. This is also confirmed for protein intake by analysis D2. However, the sensitivity analysis concerning the missing value strategy leaves some doubt on the estimated C III effects.

We agree with Bender et al. (2018b) that the most reasonable missing value strategy is to assume incomplete protocol patients belong to C III intake. Discharge from ICU is often associated with a good prognosis for patients where normal feeding is more likely to be possible. Of course, the problem of missing protocols – and how we deal with it – is potentially prone to reverse causality. By neglecting patients with incomplete protocols, we aim to handle this potential reverse causality: The estimated effects remain rather stable, though.

In total, the sensitivity analysis affirms the results from the model. However, the effect strength of moving from a medium-caloric diet to a hypocaloric diet is questionable.

7.4 Clinical inference

This section can be seen as a summary of the practical implications of our analysis. For the clinical inference, we only report the results of the cause-specific hazards model, being more likely to feature causal effects.

This analysis confirms the major findings of Bender et al. (2018b). As our main identification strategy, we use a cause-specific instead of a subdistribution hazards model. Medium-caloric diets reduce the hazard of dying in the hospital significantly compared to hypocaloric ones. The hazard does, however, not further decrease for higher-caloric diets compared to medium-caloric ones. Additionally, we find that the "hazard" to be discharged is the highest for medium-caloric diets. That means both, a fast discharge and a long survival are facilitated by medium-caloric diets. (Footnote: Between 30 and 70 % of prescribed calories or oral intake only or between 0 and 30 percent of prescribed calories and additional oral intake.)

Additionally, this analysis finds that medium-protein and high-protein diets dominate low-protein diets. Both dominant diets result in a significantly reduced hazard of dying in the hospital. The estimated hazard difference between medium and high protein diets is not significant, though. For an increased hazard of being discharged the data suggests a protein diet that starts with medium levels of protein (0.6 - 1.2 g per Kg body weight on the first few days) and increases then to high protein levels (more than 1.2 g per Kg bodyweight). This diet is also (weakly) dominant concerning the hazard of dying. Hence, this diet facilitates both, fast discharge and extended survival of patients.

The effects of caloric intake are much stronger than these for protein intake. However, both effects are significant. Also, they can be easily connected: Evaluating our approach, a medium-caloric and medium protein diet with increased levels of protein after a few days seems a dominant nutrition strategy. Interestingly, the derived recommendation is very similar to Choban et al. (2013) who recommend medium levels of calories with high levels of protein at the same time for obese critically ill ICU patients.

7.5 Related literature

As outlined in Hartl et al. (2019) the whole topic is very controversial. Guidance on how to feed on the ICU is very diverse. Even official guidelines differ substantially as outlined in Patel et al. (2017) w.r.t different countries. Thus, it is hard to identify a strong consent especially on the short-run administration of nutrition. Furthermore, most research is only occupied with explaining hospital mortality. Hence, nutrition effects on the discharge are neglected in this section.

Contradicting the findings of this thesis, recent randomised control trials (e.g. Casaer et al. (2011), Rice et al. (2012), Preiser et al. (2015)) suggest a decreased short-run caloric intake. Hartl et al. (2019) explains, though, that some of recent RCTs only have very limited external validity.

We find that our conclusions are very similar to observational studies in a similar setting. Typically, studies focus either on caloric or protein intake. Next to Bender et al. (2018b), Krishnan et al. (2003) also comes to the same conclusion as this thesis: A medium caloric (30-70 percent tolerated calories) diet is associated with optimal medical outcomes. Arabi et al. (2010) agree on this and even find adverse effects of near-target nutrition intake. Typically, many observational studies agree on low-nutrition diets being inferior w.r.t. medical outcomes compared to medium-nutrition and high-nutrition diets. For example Alberda et al. (2009) and Heyland et al. (2011) agree that high-nutrition intake is associated with the highest survival rates. Heyland et al. (2011) argue that most of these findings originate from different methodologies and outline that they find the model specification to heavily steer these findings. As other studies do not account for the complex ELRA, the approach of Bender et al. (2018b) and in this thesis actually seems advantageous. Also, we observe that many studies only use logistic regression as identification strategy with a linear effect of nutrition. The non-linear ELRA in a survival model seems the most elaborate and adequate specification. While most literature is focused on the effect of calories on hospital mortality Weijs et al. (2019) investigates protein intake in a similar setting as we do. They find that increased protein intake tends to decrease mortality significantly.

In total, most observational studies agree that either medium-nutrition or high-nutrition diets are associated with desired medical outcomes. However, many studies treat nutrition and calories identically which is strongly opposed by Hoffer (2016). Nevertheless, having studied both, caloric and protein intake, our thesis aligns with this agreement with a stronger favourability of medium-nutrition diets, especially in the short-run.

Hoffer (2016) argues, though, that from a physiological perspective the findings of the previously mentioned RCTs are plausible and seemingly reject the idea that high(er) levels of calories are helpful to boost the healing process in individuals. He comes to the conclusion that less calories in ICUs are preferable. According to him, this is not true for protein intake, though. He proposes that high levels of protein paired with a hypocaloric diet may result in desired medical outcomes. However, this idea is not validated by the RCTs carried out. Thus, Hoffer (2016) demands that in future research the emphasis should be on protein not calorie administration. Taking this idea into account, one may conclude that the effects inferred from observational studies may be interpreted in a misleading fashion. Due to the high correlation of protein and caloric intake, the effects seem to be hard to isolate. For our approach this means that it would be interesting to study caloric and protein intake jointly. In this scenario nutrition covariates would need to reflect the proposed combined diet directly. This, however, leads to a more complex model and hence more complex interpretation.

7.6 Discussion & suggestions for future research

This analysis is an update of Bender et al. (2018b) which investigates the complex ELRA in a competing risks framework. The principle identification strategy is the same as in Bender et al. (2018b). Thus, the general potential shortcomings of the analysis also apply to this study. We only discuss the most controversial one on this section in-depth:

The descriptive analysis made clear that in some cases the exogeneity of the nutrition covariates is not given. Patients with higher ApacheII scores are more likely to receive low-protein and hypocaloric nutrition on their first day in the ICU. Furthermore, the covariate of the actual intake is potentially critical. In some cases, a patient is not able to tolerate the administered intake. Subsequently, there are natural constraints on protein and caloric intake.

This is not the majority of cases and we control for the ApacheII score (as an initial assessment of the patient's health). Still, our approach is prone to a reverted causality problem. A life-threatening condition (with an a priori shorter life expectancy) may require the interruption of artificial nutrition and not the other way round. The interpretation of the cumulative effects in this thesis depends on this reverse causality problem to be rather small. Bender et al. (2018b) suggest a minimum of 4 days lag for the nutritional covariates. They do so to separate acute conditions from the nutrition effects. If a serious condition leads to death and a changed diet at the same time, the nutrition effect is not affected by this.

We show that the handling of missing patient days heavily steers the results. Nevertheless, the main specification results in very similar results as simply neglecting patients with incomplete protocols. Thus, we conclude that the results should be robust to this potentially false assumption.

Another assumption of Bender et al. (2018b) has been that individuals discharged from the hospital survived at least 30 days. We investigate the effect of loosening this assumption explicitly and find that the model results are somewhat affected.

For future research, we suggest – next to the propositions in Bender et al. (2018b) – that the lag-lead window should be estimated instead of the current approach where the lag-lead window is a priori fixed. Furthermore, it is reasonable to assume that the lag-lead windows may differ for different risks. The current approach of estimating cumulative effects would only allow dealing with the lag-lead-window as a hyperparameter to be tuned. However, hyperparameter tuning would be very expensive for the complex models presented here.

In our analysis, we model the effects of protein and calorie intake separately. This is because the two covariates are somewhat colinear. The discussion of Hoffer (2016) made clear that this may be problematic, though. In a future analysis, one could try to map the covariates in a way that they can be modeled in a joint model.

This thesis could derive a clear hypothesis of what might be a preferable diet in ICUs. Still, the evaluation of our models only made use of a very small subset of all possible diets; only 6 out of 1331 possible comparisons are conducted. Thus, this hypothesis still could be validated against the richness of different diets. While computational efforts would be relatively low for this procedure, it requires significant human evaluation effort.

Chapter 8

Conclusion

These last sections of this thesis will review the presented novelties and results. Furthermore, it represents an outlook.

8.1 Summary

This thesis is an exhaustive review of the modeling of competing risks with piece-wise exponential additive mixed models (PAMMs). It reviews concepts on PAMMs and competing risks separately. Then, it combines these concepts and proposes new models. Furthermore, this thesis applies new models to a real-world data problem. First, it assesses the problem in Bender et al. (2018b). Then, it re-evaluates their research question with new data and these new models.

8.1.1 Novelties

we transfer the cause-specific hazards and the Fine and Gray subdistribution hazards model to the PAMM context. we manage to construct analoga for the cause-specific Cox model and the subdistribution hazards Cox model. While the cause-specific PAMM is perfectly equivalent to the cause-specific Cox model, the subdistribution hazards Cox model can only be replicated. Fine and Gray (1999) use a weighted likelihood approach where the weights cannot be translated directly to the PEM setting. The weights are based on the estimated survival function of the censorship distribution. In our approach, we also use survival estimates from the censorship distribution. Instead of using weights, we directly simulate survival times from a semi-parametrically estimated distribution. These simulated survival times are used to carry-on the risk set for an appropriate time. The subdistribution hazards model is associated with missing data problems concerning time-dependent covariates In this thesis, we are agnostic to these problems as we only implement the analogon of the model proposed by Fine and Gray (1999).

we implement the proposed models in `R` and show that they work accordingly. Modeling PAMMs is mainly associated with appropriate preprocessing. Preprocessing functions for the cause-specific hazards model and all three sub-models of the subdistribution hazards model Benchmarked with the standard survival models for competing risks, we outline that our models can recover these.

Nevertheless, our thesis does not analyse the new models in an exhaustive simulation study. Additional to the modeling procedure itself, we present postprocessing functions that assist in creating estimates from the model such as cumulative hazards or cumulative incidence functions.

8.1.2 Data analysis

we use data of approximately 16000 ICU patients with information on their diet and additional health record-related data to model survival and discharge times. we loosen most modeling assumptions made by Bender et al. (2018b) who previously modeled the identical data. Our findings are in line with Bender et al. (2018b) and we find that misspecifications w.r.t. to the assumptions of Bender et al. (2018b) do not result in significantly different results.

we find that a diet which administers 0.6 g to 1.2 g protein per Kg body weight in the first days after admission is associated with desired outcomes in patients. A lifting of the protein intake to levels higher than 1.2 g for later days after admission increases these effects. The diet just described results in a decreased hospital stay and increased survival time of patients. Our findings mostly agree with similar observational studies but disagree with randomised control trials in the domain. However, according to Hoffer (2016), this disagreement could be reasoned by the fact that recent RCTs mostly focus on caloric intake and neglect protein intake which is, according to him, expected to affect medical outcomes desirably. Due to the collinearity of protein and calories, this effect may be hidden especially in observational studies.

8.2 Outlook

We embedded our contributions to the `pammtools` package into a pull request. In this thesis, we use the functions as scripts. This is because we arranged with Bender and Scheipl (2018) that we might change the front-end of our functions before finally incorporating them into `pammtools`. The functions presented in this thesis can be seen as prototypes. The final implementation into the package will follow shortly after the publishing of this thesis. From a theoretical point of view, there are three topics that we could not exclusively cover in this thesis because they are far beyond the scope of a master thesis. Nevertheless, we leave these open to future research.

First, the missing value strategy for time-dependent covariates is an interesting problem. It is a very model agnostic problem and not necessarily tied to survival analysis. Basically, one has to deal with missing data in a longitudinal context.

Second, a more exhaustive simulation study of the newly presented models is desirable. This is especially true for the subdistribution hazards model.

Third, a more systematic investigation of the identified effects in the data analysis of this thesis would be beneficial to validate the clinical findings proposed by this thesis. Furthermore, a detailed subgroup analysis as conducted in Hartl et al. (2019) would enhance insights on the effects.

Chapter 9

Appendix

9.1 Data sets

```
?veteran
```

```
## Veterans' Administration Lung Cancer study
##
## Description:
##
##      Randomised trial of two treatment regimens for lung cancer.  This
##      is a standard survival analysis data set.
##
## Usage:
##
##      veteran
##
## Format:
##
##        trt:       1=standard 2=test
##        celltype:  1=squamous,  2=smallcell,  3=adeno,  4=large
##        time:      survival time
##        status:    censoring status
##        karno:     Karnofsky performance score (100=good)
##        diagtime:  months from diagnosis to randomisation
##        age:       in years
##        prior:     prior therapy 0=no, 10=yes
##
## Source:
##
##      D Kalbfleisch and RL Prentice (1980), _The Statistical Analysis of
##      Failure Time Data_.  Wiley, New York.
```

```
?sir.adm
```

```
## Pneumonia on admission in intenive care unit patients
##
## Description:
##
##      Pneumonia status on admission for intensive care unit (ICU)
##      patients, a random sample from the SIR-3 study.
##
## Usage:
##
##      data(sir.adm)
##
## Format:
##
##      The data contains 747 rows and 4 variables:
##
##      id: Randomly generated patient id
##
##      pneu: Pneumonia indicator. 0: No pneumonia, 1: Pneumonia
##
##      status Status indicator. 0: censored observation, 1: discharged,
##           2: dead
##
##      time: Follow-up time in day
##
##      age: Age at inclusion
##
##      sex: Sex. 'F' for female and 'M' for male
##
## Source:
##
##      Beyersmann, J., Gastmeier, P., Grundmann, H., Baerwolff, S.,
##      Geffers, C., Behnke, M., Rueden, H., and Schumacher, M. Use of
##      multistate models to assess prolongation of intensive care unit
##      stay due to nosocomial infection. _Infection Control and Hospital
##      Epidemiology_, 27:493-499, 2006.
##
## Examples:
##
##      # data set transformation
##      data(sir.adm)
##      id <- sir.adm$id
##      from <- sir.adm$pneu
##      to <- ifelse(sir.adm$status==0,"cens",sir.adm$status+1)
##      times <- sir.adm$time
##      dat.sir <- data.frame(id,from,to,time=times)
##
##      # Possible transitions
##      tra <- matrix(ncol=4,nrow=4,FALSE)
##      tra[1:2,3:4] <- TRUE
##
##      na.pneu <- mvna(dat.sir,c("0","1","2","3"),
##                      tra,"cens")
##
##      if(require("lattice")) {
##      xyplot(na.pneu,tr.choice=c("0 2","1 2","0 3","1 3"),
##             aspect=1,strip=strip.custom(bg="white",
##             factor.levels=c("No pneumonia on admission -- Discharge",
##                             "Pneumonia on admission -- Discharge",
##                             "No pneumonia on admission -- Death",
##                             "Pneumonia on admission -- Death"),
##             par.strip.text=list(cex=0.9)),
##             scales=list(alternating=1),xlab="Days",
##             ylab="Nelson-Aalen esimates")
##      }
```

```
?fourD
```

```
## Placebo data from the 4D study
##
## Description:
##
##      Data from the placebo group of the 4D study. This study aimed at
##      comparing atorvastatin to placebo for patients with type 2
##      diabetes and receiving hemodialysis in terms of cariovascular
##      events. The primary endpoint was a composite of death from cardiac
##      causes, stroke and non-fatal myocardial infarction.  Competing
##      event was death from other causes.
##
## Usage:
##
##      data(fourD)
##
## Format:
##
##      A data frame with 636 observations on the following 7 variables.
##
##      'id' Patients' id number
##
##      'sex' Patients' gender
##
##      'age' Patients' age
##
##      'medication' Character vector indicating treatment affiliation.
##           Here only equal to '"Placebo"'
##
##      'status' Status at the end of the follow-up. 1 for the event of
##           interest, 2 for death from other causes and 0 for censored
##           observations
## 
##      'time' Survival time
## 
##      'treated' Numeric vector indicated whether patients are treated or
##           not. Here always equal to zero
## 
## Source:
## 
##      Wanner, C., Krane, V., Maerz, W., Olschewski, M., Mann, J., Ruf,
##      G., Ritz, E (2005). Atorvastatin in patients with type 2 diabetes
##      mellitus undergoing hemodialysis. New England Journal of Medicine,
##      353(3), 238-248.
## 
## References:
## 
##      Allignol, A., Schumacher, M., Wanner, C., Dreschler, C. and
##      Beyersmann, J. (2010). Understanding competing risks: a simulation
##      point of view. Research report.
## 
## Examples:
## 
##      data(fourD)
```

```
?pbc
```

```
## Mayo Clinic Primary Biliary Cirrhosis Data
## 
## Description:
## 
##      D This data is from the Mayo Clinic trial in primary biliary
##      cirrhosis (PBC) of the liver conducted between 1974 and 1984.  A
##      total of 424 PBC patients, referred to Mayo Clinic during that
##      ten-year interval, met eligibility criteria for the randomized
##      placebo controlled trial of the drug D-penicillamine.  The first
##      312 cases in the data set participated in the randomized trial and
##      contain largely complete data.  The additional 112 cases did not
##      participate in the clinical trial, but consented to have basic
##      measurements recorded and to be followed for survival.  Six of
##      those cases were lost to follow-up shortly after diagnosis, so the
##      data here are on an additional 106 cases as well as the 312
##      randomized participants.
## 
##      A nearly identical data set found in appendix D of Fleming and
##      Harrington; this version has fewer missing values.
## 
## Usage:
## 
```

```
##      pbc
##
## Format:
##
##        age:       in years
##        albumin:   serum albumin (g/dl)
##        alk.phos:  alkaline phosphotase (U/liter)
##        ascites:   presence of ascites
##        ast:       aspartate aminotransferase, once called SGOT (U/ml)
##        bili:      serum bilirunbin (mg/dl)
##        chol:      serum cholesterol (mg/dl)
##        copper:    urine copper (ug/day)
##        edema:     0 no edema, 0.5 untreated or successfully treated
##                   1 edema despite diuretic therapy
##        hepato:    presence of hepatomegaly or enlarged liver
##        id:        case number
##        platelet:  platelet count
##        protime:   standardised blood clotting time
##        sex:       m/f
##        spiders:   blood vessel malformations in the skin
##        stage:     histologic stage of disease (needs biopsy)
##        status:    status at endpoint, 0/1/2 for censored, transplant, dead
##        time:      number of days between registration and the earlier of death,
##                   transplantion, or study analysis in July, 1986
##        trt:       1/2/NA for D-penicillmain, placebo, not randomised
##        trig:      triglycerides (mg/dl)
##
## Source:
##
##      T Therneau and P Grambsch (2000), _Modeling Survival Data:
##      Extending the Cox Model_, Springer-Verlag, New York.  ISBN:
##      0-387-98784-3.
##
## See Also:
##
##      'pbcseq'
```

9.2 Competing risks for PAMMs

The following section refers to chapter 5

9.2.1 Missing values

As outlined throughout this thesis, a core problem to be solved when modeling subdistribution hazards are missing time-dependent covariates for competing risks where the hypothetical censorship time is larger than the actual one.

Our solution is a simple carry-on. We use the `pbc` data.

```r
data("pbc", package = "survival")
pbc <- pbc %>% mutate(bili = log(bili), protime = log(protime))
pbcseq <- pbcseq %>% mutate(bili = log(bili), protime = log(protime))
pbc <- pbc %>% filter(id <= 312) %>%
  select(id:sex, bili, protime)

## make sure status only in pbc
pbcseq <- pbcseq %>% select(- "status")

pbc_ped <- as_ped_cr_sh(
  data = list(pbc, pbcseq),
  formula = Surv(time, status) ~ . + concurrent(bili, protime, tz_var = "day"),
  id = "id")
data.frame(time = pbcseq$day, protime = pbcseq$protime)[1:2, ]
```

time	protime
0	2.501
192	2.416

```r
data.frame(time = pbc_ped$`1`$tend, protime = pbc_ped$`1`$protime)[39:46, ]
```

	time	protime
39	188	2.501
40	189	2.501
41	190	2.501
42	191	2.501
43	192	2.501
44	193	2.416
45	194	2.416
46	195	2.416

For cumulative effects the default in `pammtools` is different as we can see if we modify our example from chapter 2 slightly:

```r
data_4 <- data.frame(id = 1:4,
                     obs_times = c(3.5, 2.5, 5, 7),
                     status = c(0, 1, 1, 0))
data_4_cumu <- data.frame(id = c(1, 1, 1,
                                      2, 2, 2,
                                      3, 3, 3,
                                      4, 4, 4, 4),
                               day = c(rep(1:3, 3), 1, 2, 4, 5),
                               x1 = c(0, 1, 1,
                                      0, 1, 0,
                                      1, 0, 3,
                                      1, 5, 1, 1))

data_4
```

id	obs_times	status
1	3.5	0
2	2.5	1
3	5.0	1
4	7.0	0

```
data_4_cumu
```

id	day	x1
1	1	0
1	2	1
1	3	1
2	1	0
2	2	1
2	3	0
3	1	1
3	2	0
3	3	3
4	1	1
4	2	5
4	4	1
4	5	1

Now, we have one additional individual with a survival time longer than the other and also more days observed (`id 4`). Furthermore, for this individual, there is no measurement on day `3` but on day `4`. Furthermore, there is one individual who also lived longer than we have exposure data (`id 3`).

```
ped_4 <- as_ped(
  data    = list(data_4, data_4_cumu),
  formula = Surv(obs_times, status) ~ . + cumulative(day, x1, tz_var = "day"),
  cut     = seq(0, max(data_4$obs_times), 0.5),
  id = "id")

data.frame(id = ped_4$id, ped_4$day)[13:22, ] # for id 3
```

	id	day1	day2	day3	day4
13	3	1	2	3	0
14	3	1	2	3	0
15	3	1	2	3	0
16	3	1	2	3	0
17	3	1	2	3	0
18	3	1	2	3	0
19	3	1	2	3	0
20	3	1	2	3	0
21	3	1	2	3	0
22	3	1	2	3	0

```r
data.frame(id = ped_4$id, ped_4$x1)[13:22, ]
```

	id	x11	x12	x13	x14
13	3	1	0	3	0
14	3	1	0	3	0
15	3	1	0	3	0
16	3	1	0	3	0
17	3	1	0	3	0
18	3	1	0	3	0
19	3	1	0	3	0
20	3	1	0	3	0
21	3	1	0	3	0
22	3	1	0	3	0

```r
data.frame(id = ped_4$id, LL = ped_4$LL)[13:22, ]
```

	id	LL.1	LL.2	LL.3	LL.4	LL.5
13	3	0	0	0	0	0
14	3	0	0	0	0	0
15	3	1	0	0	0	0
16	3	1	0	0	0	0
17	3	1	1	0	0	0
18	3	1	1	0	0	0
19	3	1	1	1	0	0
20	3	1	1	1	0	0
21	3	1	1	1	1	0
22	3	1	1	1	1	0

```r
data.frame(id = ped_4$id, ped_4$day)[23:36, ] # for id 4
```

	id	day1	day2	day3	day4
23	4	1	2	4	5
24	4	1	2	4	5
25	4	1	2	4	5
26	4	1	2	4	5
27	4	1	2	4	5
28	4	1	2	4	5
29	4	1	2	4	5
30	4	1	2	4	5
31	4	1	2	4	5
32	4	1	2	4	5
33	4	1	2	4	5
34	4	1	2	4	5
35	4	1	2	4	5
36	4	1	2	4	5

```r
data.frame(id = ped_4$id, ped_4$x1)[13:36, ]
```

	id	x11	x12	x13	x14
13	3	1	0	3	0
14	3	1	0	3	0
15	3	1	0	3	0
16	3	1	0	3	0
17	3	1	0	3	0
18	3	1	0	3	0
19	3	1	0	3	0
20	3	1	0	3	0
21	3	1	0	3	0
22	3	1	0	3	0
23	4	1	5	1	1
24	4	1	5	1	1
25	4	1	5	1	1
26	4	1	5	1	1
27	4	1	5	1	1
28	4	1	5	1	1
29	4	1	5	1	1
30	4	1	5	1	1
31	4	1	5	1	1
32	4	1	5	1	1
33	4	1	5	1	1
34	4	1	5	1	1
35	4	1	5	1	1
36	4	1	5	1	1

```r
data.frame(id = ped_4$id, LL = ped_4$LL)[13:36, ]
```

	id	LL.1	LL.2	LL.3	LL.4	LL.5
13	3	0	0	0	0	0
14	3	0	0	0	0	0
15	3	1	0	0	0	0
16	3	1	0	0	0	0
17	3	1	1	0	0	0
18	3	1	1	0	0	0
19	3	1	1	1	0	0
20	3	1	1	1	0	0
21	3	1	1	1	1	0
22	3	1	1	1	1	0
23	4	0	0	0	0	0
24	4	0	0	0	0	0
25	4	1	0	0	0	0
26	4	1	0	0	0	0
27	4	1	1	0	0	0
28	4	1	1	0	0	0
29	4	1	1	1	0	0
30	4	1	1	1	0	0
31	4	1	1	1	1	0
32	4	1	1	1	1	0
33	4	1	1	1	1	1
34	4	1	1	1	1	1
35	4	1	1	1	1	1
36	4	1	1	1	1	1

The lag-lead window is simply carried on. However, `x1` and `day` are set to zero when the observed exposure is already over. Days without exposure are “skipped”. This eventually means that the covariate is set to zero on days that are not observed. That means that the cumulative covariate is set to carry-on the cumulative exposure until the end of observed exposure (i.e. the latest exposure time observed). This assumption is reasonable if there is no missing data. Or in other words: every single exposure was recorded. For example, in a medical study, all patients will have very different records and there will be no missing data if treatment is administered on a non-regular basis.

However, in the subdistribution context, this is **explicitely** not the case. Still, one could use this cumulative carry-on as a way to treat missing data. On the other hand, an imputation approach or another form of carrying-on, the carrying on of the lastly measured exposure, would be feasible.

While the default approach in `pammtools` is working for ordinary time-dependent and complex cumulative effects, we think that there are more appropriate solutions that are not trivial. However, with appropriate preprocessing one can have `pammtools` impute more smartly. The follow-up time is directly linked to the observed time of the time-dependent covariates or via the lag-lead-function. Thus, `pammtools` automatically ignores irrelevant longitudinal observations. For instance, the two `ped` objects below are identical.

```r
data_3 <- data.frame(id = 1:3,
                     obs_times = c(1, 0.5, 2),
                     status = c(0, 1, 1),
                     x1 = c(0, 4, 5))
```

```
data_3_long <- data.frame(id = c(1, 1, 2, 2, 2, 2, 3, 3, 3),
                          day = c(0, 1,
                                  0, 0.2, 0.3, 0.4,
                                  0, 1, 2),
                          x2 = c(1, 1,
                                 1, 0, 2, 1,
                                 0, 1, 2))

data_3_long_2 <- data.frame(id = c(1, 1, 2, 2, 2, 2, 2, 3, 3, 3),
                          day = c(0, 1,
                                  0, 0.2, 0.3, 0.4, 1,
                                  # additional obs AFTER failure
                                  0, 1, 2),
                          x2 = c(1, 1,
                                 1, 0, 2, 1, 1,
                                 0, 1, 2))

ped_3 <- as_ped(
  data    = list(data_3, data_3_long),
  formula = Surv(obs_times, status) ~ . + concurrent(x2, tz_var = "day"),
  id      = "id",
  cut     = seq(0, max(data_3$obs_times), 0.4))
ped_3_2 <- as_ped(
  data    = list(data_3, data_3_long_2),
  formula = Surv(obs_times, status) ~ . + concurrent(x2, tz_var = "day"),
  id      = "id",
  cut     = seq(0, max(data_3$obs_times), 0.4))

all.equal(ped_3, ped_3_2)
```

```
## [1] TRUE
```

9.3 Data analysis

The following sections refer to chapter 7.

9.3.1 Data clearing

The data is supplied via a SAS database and comes in different files, one for ICU characteristics, one for patient data, and one for the nutrition protocols. The ICU data set is almost completely neglected, as later on ICUs will be included in the regression model via random effects.

The preprocessing starts with renaming the SAS covariate names to the naming scheme of Bender (2018) and is followed by the deletion of irrelevant columns. The data has many timestamps. From these one needs to infer the resulting numeric measures, like the survival time. Furthermore, these timestamps are a source to potential inconsistencies in the data. While the absolute number of cases like these is very low, data where e.g. the mechanical ventilation has been discontinued before it started, is to be seen as critical. Depending on the situation, cases are either dropped or the values are replaced by a decent guess from other timestamps or values. For instance, there are some patients for which the beginning of the administration of artificial nutrition has been recorded long before the admission to hospital. These observations were removed. Missing data is very rare and hence, we simply drop incomplete cases.

9.3.2 Data mapping

This section deals with the mapping of our data to the data of Bender et al. (2018b).

Unfortunately, the mapping of the data of Bender et al. (2018b) and ours is more difficult than anticipated. This is because the new data sets have a more coherent scheme for their IDs.

In this section, however, we map the data of Bender et al. (2018b) with our data.

Bender et al. (2018b) present their raw data in their GitHub repo not completely pre-processed. For example, the final transformation of the confounders for day 2 to day 4 has not yet taken place. Based on this transformation, there are further exclusions. Thus, there will necessarily be observations in the old data set which are not in the new one. However, we don't intend to compare data that has not been used in the analysis. Hence, for mapping, it is more relevant to find the data in the new data set in the old one than the other way round. We attached a script which assesses the mapping:

For 2007 all IDs could be eventually mapped (using some workarounds).

For 2008 IDs very easily to map; however, 197 patient days could not be mapped. This is, however a reasonably low.

For 2009, the IDs can only be mapped with extensive engineering. Even then, 1738 patient days cannot be mapped. This is a significant number. We think that by inspection of the single cases a transformation is feasible. However, this should not be necessary and hence neglected at this point.

For 2011, the same problem arises. 3608 patient days cannot be mapped eventually.

The majority of cases (89599 patient days out of 111088) could be mapped and be compared. However, the potential inconsistencies need to be investigated in detail with the data issuer.

We find for in total around 1092 patient days differ in reported calories of at least 25 kcal calories intake and 994 patient days a difference of at least 0.1 g protein intake per Kg bodyweight. If this threshold was surpassed, the average amount of difference was substantial. In the new data delivery, most of the concerned patient days were reported with 0 kcal and 0 g protein, while in the previous data there has been reported intake.

The protocols only differ on the last day, though. This is due to the different handling of incomplete protocols. we impute incomplete last days on the ICU **and** incomplete protocols due to censoring. Our imputation is, however, only base on the less noisy categorical features `calCat2` and `calCat3` (and `proteinCat2` and `proteinCat3`, respectively). Our main specification is setting the categories to a C III intake (this data has been used for mapping). we also investigate random sampling of the category and a carry-on of every different category.

Chapter 10

Declaration of originality

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations. I have clearly referenced in accordance with departmental requirements. Additionally, I confirm that this work has not been previously or concurrently used for other courses. I confirm that I understand that my work may be electronically checked for plagiarism and appreciate that any false claim in respect of this work will result in disciplinary action.

Philipp Kopper, Munich, April 28, 2020

Bibliography

Aalen, O. O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4):177–190.

Alberda, C., Gramlich, L., Jones, N., Jeejeebhoy, K., Day, A. G., Dhaliwal, R., and Heyland, D. K. (2009). The relationship between nutritional intake and clinical outcomes in critically ill patients: results of an international multicenter observational study. *Intensive care medicine*, 35(10):1728–1737.

Arabi, Y. M., Haddad, S. H., Tamim, H. M., Rishu, A. H., Sakkijha, M. H., Kahoul, S. H., and Britts, R. J. (2010). Near-target caloric intake in critically ill medical-surgical patients is associated with adverse outcomes. *Journal of Parenteral and Enteral Nutrition*, 34(3):280–288.

Austin, P. C. and Fine, J. P. (2017). Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in medicine*, 36(8):1203–1209.

Austin, P. C., Latouche, A., and Fine, J. P. (2020). A review of the use of time-varying covariates in the fine-gray subdistribution hazard competing risk regression model. *Statistics in medicine*, 39(2):103–113.

Belot, A., Abrahamowicz, M., Remontet, L., and Giorgi, R. (2010). Flexible modeling of competing risks in survival analysis. *Statistics in medicine*, 29(23):2453–2468.

Bender, A. (2018). *Flexible modeling of time-to-event data and exposure-lag-response associations.* PhD thesis, lmu.

Bender, A., Groll, A., and Scheipl, F. (2018a). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321.

Bender, A. and Scheipl, F. (2018). pammtools: Piece-wise exponential additive mixed modeling tools. *arXiv preprint arXiv:1806.01042*.

Bender, A., Scheipl, F., Hartl, W., Day, A. G., and Küchenhoff, H. (2018b). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*, 20(2):315–331.

Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing risks and multistate models with R.* Springer Science & Business Media.

Casaer, M. P., Mesotten, D., Hermans, G., Wouters, P. J., Schetz, M., Meyfroidt, G., Van Cromphaut, S., Ingels, C., Meersseman, P., Muller, J., et al. (2011). Early versus late parenteral nutrition in critically ill adults. *New England Journal of Medicine*, 365(6):506–517.

Cederholm, T., Jägren, C., and Hellström, K. (1995). Outcome of protein-energy malnutrition in elderly medical patients. *The American journal of medicine*, 98(1):67–74.

Choban, P., Dickerson, R., Malone, A., Worthington, P., Compher, C., for Parenteral, A. S., and Nutrition, E. (2013). Aspen clinical guidelines: nutrition support of hospitalized adult patients with obesity. *Journal of Parenteral and Enteral nutrition*, 37(6):714–744.

Clark, M. (2019). *Generalized additive models.*

Cox, D. (1972). Regression models and life tables (with discussion). lr statist.

Danieli, C. and Abrahamowicz, M. (2019). Competing risks modeling of cumulative effects of time-varying drug exposures. *Statistical methods in medical research*, 28(1):248–262.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509.

Friedman, M. et al. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1):101–113.

Gasparrini, A. (2014). Modeling exposure–lag–response associations with distributed lag non-linear models. *Statistics in medicine*, 33(5):881–899.

Haller, B. (2014). *The analysis of competing risks data with a focus on estimation of cause-specific and subdistribution hazard ratios from a mixture model.* PhD thesis, lmu.

Hartl, W. H., Bender, A., Scheipl, F., Kuppinger, D., Day, A. G., and Küchenhoff, H. (2019). Calorie intake and short-term survival of critically ill patients. *Clinical Nutrition*, 38(2):660–667.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Heyland, D. K., Cahill, N., and Day, A. G. (2011). Optimal amount of calories for critically ill patients: depends on how you slice the cake! *Critical care medicine*, 39(12):2619–2626.

Hoffer, L. J. (2016). Protein requirement in critical illness. *Applied Physiology, Nutrition, and Metabolism*, 41(5):573–576.

Iljon, T. (2013). *Calculating confidence intervals for the cumulative incidence function while accounting for competing risks: comparing the Kalbfleisch-Prentice method and the Counting Process method.* PhD thesis.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G., Sirio, C. A., Murphy, D. J., Lotring, T., Damiano, A., et al. (1991). The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. *Chest*, 100(6):1619–1636.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis.* CRC Press.

Krishnan, J. A., Parce, P. B., Martinez, A., Diette, G. B., and Brower, R. G. (2003). Caloric intake in medical icu patients: consistency of care with guidelines and relationship to clinical outcomes. *Chest*, 124(1):297–305.

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American journal of epidemiology*, 170(2):244–256.

Lunn, M. and McNeil, D. (1995). Applying cox regression to competing risks. *Biometrics*, pages 524–532.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

McCulloch, C. E. and Neuhaus, J. M. (2005). Generalized linear mixed models. *Encyclopedia of biostatistics*, 4.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Patel, J. J., Lemieux, M., McClave, S. A., Martindale, R. G., Hurt, R. T., and Heyland, D. K. (2017). Critical care nutrition support best practices: Key differences between canadian and american guidelines. *Nutrition in Clinical Practice*, 32(5):633–644.

Preiser, J.-C., van Zanten, A. R., Berger, M. M., Biolo, G., Casaer, M. P., Doig, G. S., Griffiths, R. D., Heyland, D. K., Hiesmayr, M., Iapichino, G., et al. (2015). Metabolic and nutritional support of critically ill patients: consensus and controversies. *Critical care*, 19(1):35.

Rice, T., Wheeler, A., Thompson, B., Steingrub, J., Hite, R., Moss, M., Morris, A., Dong, N., and Rock, P. (2012). National heart, lung, and blood institute acute respiratory distress syndrome (ards) clinical trials network. initial trophic vs full enteral feeding in patients with acute lung injury: the eden randomized trial. *JAMA*, 307(8):795–803.

Rotnitzky, A. G. and Robins, J. (2005). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*, 4.

Sargent, D. J. (1998). A general framework for random effects survival analysis in the cox proportional hazards setting. *Biometrics*, pages 1486–1497.

Sharma, K., Mogensen, K. M., and Robinson, M. K. (2019). Pathophysiology of critical illness and role of nutrition. *Nutrition in Clinical Practice*, 34(1):12–22.

Singer, P., Blaser, A. R., Berger, M. M., Alhazzani, W., Calder, P. C., Casaer, M. P., Hiesmayr, M., Mayer, K., Montejo, J. C., Pichard, C., et al. (2019). Espen guideline on clinical nutrition in the intensive care unit. *Clinical nutrition*, 38(1):48–79.

Sylvestre, M.-P. and Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in medicine*, 28(27):3437–3453.

Szakmany, T., Walters, A. M., Pugh, R., Battle, C., Berridge, D. M., and Lyons, R. A. (2019). Risk factors for 1-year mortality and hospital utilization patterns in critical care survivors: A retrospective, observational, population-based data linkage study. *Critical care medicine*, 47(1):15–22.

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

Weijs, P. J., Mogensen, K. M., Rawn, J. D., and Christopher, K. B. (2019). Protein intake, nutritional status and outcomes in icu survivors: a single center cohort study. *Journal of clinical medicine*, 8(1):43.

Wood, S. N. (2001). mgcv: Gams and generalized ridge regression for r. *R news*, 1(2):20–25.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Xie, Y. (2016). *Bookdown: Authoring books and technical documents with R markdown*. CRC Press.