Structured Fusion Lasso Penalised Multi-state Models

Holger Sennhenn-Reulen[a,b,c] Thomas Kneib[a]

[a] Chair of Statistics, University of Göttingen, Germany.

[b] Leibniz-ScienceCampus Primate Cognition, German Primate Center, Leibniz Institute for Primate Research, Göttingen, Germany.

[c] Cognitive Ethology Laboratory, German Primate Center, Leibniz Institute for Primate Research, Göttingen, Germany.

Corresponding author: Holger Sennhenn-Reulen, Cognitive Ethology Laboratory, German Primate Center, Leibniz Institute for Primate Research, Göttingen, Germany. E-mail: hsennhenn-reulen@dpz.eu.

Abstract: Multi-state models generalize survival or duration time analysis to the estimation of transition-specific hazard rate functions for multiple transitions. When each of the transition-specific risk functions is parametrized with several distinct covariate effect coefficients, this leads to a model of potentially high dimension. To decrease the parameter space dimensionality and to work out a clear image of the underlying multi-state model structure, one can either aim at setting some coefficients to zero or to make coefficients for the same covariate but two different transitions equal. The first issue can be approached by penalising the absolute values of the covariate coefficients as in lasso regularisation. If, instead, absolute differences between coefficients of the same covariate on different transitions are penalized, this leads to sparse competing risk relations within a multi-state model, i.e. equality of covariate effect coefficients. In this paper, a new estimation approach providing sparse multi-state modelling by the above principles is established, based on the estimation of multi-state models and a simultaneous penalisation of the L_1-norm of covariate coefficients and their differences in a structured way. The new multi-state modelling approach is illustrated on peritoneal dialysis study data and implemented in the R package `penMSM`.

Keywords: Multi-state models; Regularisation; Structured fusion Lasso penalty; Cross-transition effects.

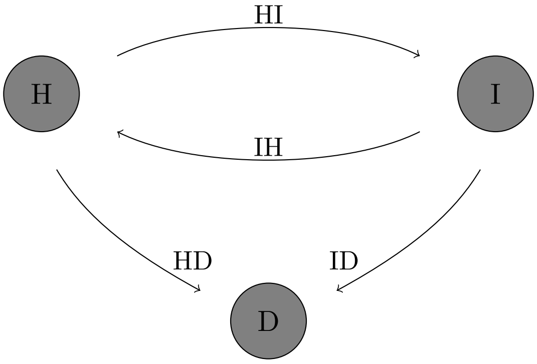

Figure 1: State-chart of an illness-death model with recovery illustrating the underlying process that leads to the sequences of events. State H denotes healthy, state I denotes illness, and state D denotes death. Transition IH is the representative for recovery.

1. Introduction

Multi-state models are a general model class for analysing the timing of events with a wide range of applications in medicine. In a general definition, a multi-state model is characterized as a system of multivariate survival data where the individuals under study may experience a sequence of transitions across time. Each transition is characterized by an entry and an exit state, the time when the entry state is reached and the duration of the sojourn time until the transition is either observed or censored. The durations of the sojourn times are then influenced by transition-specific covariate effects. The most prominent model for encompassing such covariate effects is the Cox proportional hazards model [1]:

$$\lambda_{q,i}\left(t\right) = \lambda_{q,0}\left(t\right) \cdot \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}_q\right),$$

with transition set $q \in \{1,\ldots,Q\} = \mathcal{Q}$, individuals $i = 1,\ldots,N$, time t, transition-specific baseline hazard rate function $\lambda_{q,0}\left(t\right)$, individual-specific covariate vectors $\mathbf{x}_i$ as a collection of covariate observations $x_{p,i}$, $p = 1,\ldots,P$, and corresponding transition-specific covariate coefficient vectors $\boldsymbol{\beta}_q$. The product $\mathbf{x}_i^\top \boldsymbol{\beta}_q$ results in the individual- and transition-specific linear predictor $\eta_{q,i}$.

Each transition sequence is characterized by a series of distinct entry and exit states following paths of possible transitions. This system of paths can be illustrated by a state-chart, where distinct states are treated as nodes and possible transitions are represented by directed arrows. In general, transitions between two states may be reversible or irreversible: in the first case, only one arrow exists between the two states while in the second case two arrows connect the two states. Figure 1 shows the state-chart for an illness-death model with recovery. This is a three-state model with the transitions between the states healthy (H) and illness (I) being reversible, while the transitions to death (D) are considered as being irreversible.

The transitions of a multi-state model are categorically scaled characteristics and some of them may have a closer relation to each other with respect to their practical interpretation. In the illness-death model, the transitions that point from H to I (HI) or from I to D (ID) have a major common attribute: they both lead to the aggravation of a patients health situation. While some of the risk factors that are associated with the sojourn times in the entry states of these transitions may have different effects strengths, there may be others that can be described with the same effect magnitude on both transitions, i.e. $\beta_{\mathrm{HI}} = \beta_{\mathrm{ID}} \neq \beta_{\mathrm{IH}}$ for covariate x_p and transition I to H (IH). In the underlying data generating process, this case of equal effects is mainly controlled by the aggravation component, while the signal of the single transition-specific components is negligible in terms of variable selection.
Covariate effects that are equal across two, or – in other situations – more transitions will be denoted as *cross-transition effects* in the remainder of this article. Following ideas from Thall and Lachin [2], the value of cross-transition effects with respect to the interpretation of results gets clearest when we consider the relative transition hazard rate function for two transitions q, q', where both baseline hazard rates and all covariate effects are unique:

$$\frac{\lambda_{q,i}\left(t\right)}{\lambda_{q',i}\left(t\right)} = \frac{\lambda_{q,0}\left(t\right) \cdot \exp\left(x_{1,i}\beta_{1.q}\right) \cdot \ldots \cdot \exp\left(x_{P,i}\beta_{P.q}\right)}{\lambda_{q',0}\left(t\right) \cdot \exp\left(x_{1,i}\beta_{1.q'}\right) \cdot \ldots \cdot \exp\left(x_{P,i}\beta_{P.q'}\right)}.$$

If $\beta_{p.q} = \beta_{p.q'}$, the respective term will cancel out for any value of $x_{p,i}$. If $\lambda_{q,0}\left(t\right) = \gamma_{q.q'}\lambda_{q',0}\left(t\right)$, the baseline hazards are proportional and the relative transition hazard rate function simplifies to:

$$\gamma_{q.q'}\frac{\exp\left(x_{1,i}\beta_{1.q}\right) \cdot \ldots \cdot \exp\left(x_{P,i}\beta_{P.q}\right)}{\exp\left(x_{1,i}\beta_{1.q'}\right) \cdot \ldots \cdot \exp\left(x_{P,i}\beta_{P.q'}\right)}.$$

Analyses of this type are feasible using the piecewise exponential model approach. For partial likelihood analyses, the latter investigation of proportionality of baseline hazard rates is not directly accessible during the modelling stage.
Another type of grouping structure for transitions is present in the five-state model which will be analysed during this article with the purpose to illustrate the established multi-state modelling approach. In this data set, patients with a chronic kidney disease participated in a peritoneal dialysis program at the Peritoneal Dialysis Unit, Nephrology Department, Hospital Geral de Santo António, Porto, Portugal, between 1980 and 2011. All of the 425patients under study start with the entrance (E) to the peritoneal dialysis program and are at-risk for the transition into the transient state peritonitis (P), or into one of three absorbing states: death (D), transfer to haemodialysis (H), and renal transplantation (R). If a patient has reached the transient state P, she or he is immediately at risk again to reach one of the three absorbing states D, H, or R. The state-chart for this multi-state model is presented in Figure 2. This data is originally analysed in an un-penalised multi-state model by Laetitia Teixeira, Anabela Rodrigues, and Denisa Mendonça from the University of Porto, Portugal, and Carmen Cadarso-Suárez from the University of Santiago de Compostela, Spain, (Unpublished)

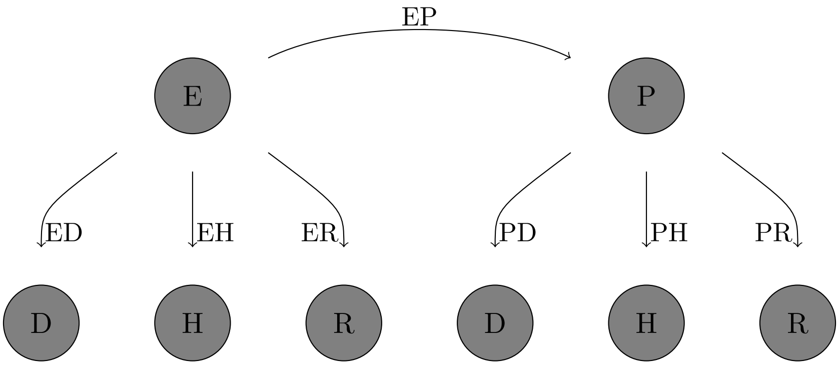

Figure 2: State-chart for the multi-state model on the peritoneal dialysis program data. State E denotes the entrance to the peritoneal dialysis program, state P denotes an affection with peritonitis, state D denotes the death of a patient, state H denotes a transfer to haemodialysis, and R denotes a renal transplantation.

and we will refer to these results (given in Table 1) as a benchmark model in Section 3. A closer look on the results from Teixeira *et al.* as given by Table 1, suggests – in combination with the model structure illustrated in Figure 2 – to take into account the possibility of cross-transition effects: While some covariates or risk factors have considerable differences between the estimated coefficients for the same absorbing exit state, others, such as the presence of diabetes in connection with a transition to state D, exhibit only small differences. These small differences are possibly negligible from an applied perspective and a data-driven algorithm that sets these to 0 may be of a high practical value.

Several general overview articles for multi-state models face the point of shared features across transition types, e.g.: "Of course it is not of much interest to make a joint model which gives the same results as the models fitted separately to each transition. The point comes from reducing this model to one which is more parsimonious yet sensible." [3], "[...] go a step further in order to analyse more parsimonious models where some baseline intensities are proportional or where some covariates have the same effect on several transition intensities." [4], or "Interaction effects between covariates and strata may be used to assess whether covariate effects vary across competing outcomes [...]" [5]. These articles provide different types of statistical tests, but none gives a data-driven and automatic algorithm to solve this problem while answering the question about the occurrence of shared features across transition types. Schmidtmann *et al.* [6] use a boosting approach for coupled selection of covariates across transitions, but they in fact yield different estimates for each transition-specific hazard rate function. Therefore they select effects simultaneously but donot enforce group sparsity as defined in the following. In general, the number of degrees of freedom of a multi-state model, i.e. the number of regression parameters to be estimated, equals the number of all transition-specific effect parameters. This quantity increases both with the number of covariates to be considered and the number of transitions. Consequently, in a multi-state model with a large number of transitions and/or potential covariates, the maximum model is quite complex and hard to interpret. To address this difficulty and also to alleviate overfitting, a practical solution is to control the number of degrees of freedom of the model and to assume that only a few of all the transition-specific effects are actually relevant for prediction. That is, the vector of model coefficients is sparse, meaning that the transition-specific covariates whose associated model coefficients take a value equal to zero do not contribute to the decisions made by the multi-state process. If performed in a non-automated way, the search for this subset of relevant model coefficients is a very labour-intensive task. For an automated, i.e. data-driven approach, different strategies to be used for estimating coefficients under the sparsity assumption are described in the literature either for regression models such as linear or generalized linear models or for the Cox proportional hazards model in event time analysis. The Least angle shrinkage and selection operator (Lasso) [7] is one of the most prominent approaches, with regularisation of the absolute value of model coefficients as central idea. This has the consequence that single model coefficients may be estimated to be exactly equal to zero, and the corresponding covariates drop out of the model. The Lasso has been successfully applied to Cox proportional hazards models [8] and has also been generalized to penalties with additional positivity constraints, constraints on the absolute differences between model coefficients or constraints on squared model coefficient values [9]. This is accompanied with an alternative estimation algorithm using a series of directional Taylor approximations and results in the very robust performance of the R [10] add-on package `penalized` [11]. However, the implementation in `penalized` as well as the presentation by Goeman [9] allows to imply constraints on the absolute differences between model coefficients only for ordinal covariates or in feature selection. This does not allow to adequately use prior knowledge about a possible grouping structure between transitions.

The process of introducing the covariate coefficients under the sparsity assumption can be facilitated when prior information is available about groups of features that are expected to be jointly relevant or jointly irrelevant for prediction [12], i.e. when different groups of covariate coefficients are expected to be jointly equal to or jointly different from zero. Finding this type of information can be difficult in practice, but is in many practical multi-state models directly approachable from the underlying state-chart. Having this information at hand might be beneficial to improve the estimates of the covariate coefficients and to reduce the number of samples required to obtain a good generalization performance. As described by Puig *et al.* [13], there is in general a very wide range of applications where sparsity at the group level is beneficial, including regression with grouped variables, source localization, or whole genome association mapping. As with the individual sparsity assumption, sparsity at the group level can be introduced in the estimation process of the model coefficients by considering specific regularization norms at the group level.

To sum up the central points: regularization is a natural task in multi-state modelling induced by the highly parametrized nature of this model class. Using additional information about the structure of the model can be beneficial for the estimation results with respect to interpretability and generalisability. Both points can be incorporated into a single regularized estimation approach for multi-state models based on penalties for absolute values of model coefficients in combination with penalties for particular selected pairwise differences of model coefficients. The selection of pairs can be gathered from the state-chart of the multi-state model. The theory and implementation of this approach is described in the following section, based on a model with proportional hazards assumption. The theoretical concepts and practical steps are described and the performance is illustrated on a real data application.

2. Structured fusion Lasso penalised multi-state modelling

This section describes the construction of structured fusion Lasso penalised multi-state models. The basic concepts of multi-state models are very thoroughly described by Andersen and Keiding [4] and Andersen *et al.* [14], and the following introduction is based on Andersen and Keiding [4].

2.1. Basic likelihood formulation for multi-state models

The underlying mathematical concept of a multi-state models is the multi-state process $Y(t)$, $t \in \mathcal{T}$, i.e. a stochastic process with a finite state space $\mathcal{K} = \{1, \ldots, k, \ldots, k', \ldots, K\}$ and right-continuous sample path $Y(t+) = Y(t)$, where $t+$ denotes the limit from the right to t (i.e. in a mathematically imprecise, informal interpretation the time point immediately after t). The sample paths of such a process are constant between the times of transitions, with time taking on values in $\mathcal{T} = [0, t_{\max}]$, with $0 < t_{\max} < \infty$. Of course, left-truncation or right-censoring can also be present in multi-state models. Over the course of time, a multi-state process $Y(\cdot)$ generates a history $\mathcal{Y}_t$, which is the σ-algebra generated by the observed sample path in the interval $[0, t]$.
We may define transition-specific transition probabilities:

$$P_q(s,t) = P_{k.k'}(s,t) = \mathbb{P}\left(Y(t) = k' \mid Y(s) = k, \mathcal{Y}_{s-}\right),$$

for $k, k' \in \mathcal{K}$, $s, t \in \mathcal{T}$, $s \leq t$, transitions denoted by $q = k.k' \in \mathcal{Q}$ referring to transitions from k to k', $k \neq k'$, and $s-$ defined in analogy to $t+$.
Using this definition, we may furthermore make the Markov assumption to define transition-specific transition intensities:

$$\lambda_q(t) = \lim_{\Delta_t \downarrow 0} \frac{P_q(t, t+\Delta_t)}{\Delta_t},$$

which we shall assume to exist.
State-charts, such as given in Figures 1 and 2, are a useful tool for graphical representations of the transitions of a multi-state model. Let $\mathcal{Q} = \{1, \ldots, q, \ldots, q', \ldots, Q\}$ define the set of observable transitions.

Assume that multi-state processes $Y_i\left(t\right)$ are observed over intervals $[0, t_{\max,i}]$ for individuals i, $i = 1, \ldots, N$, where $t_{\max,i}$ is the time of termination of the observation for individual i. Since the individual processes are constant between observed transitions, it is equivalent to record the state at the origin $Y_i\left(0\right)$ and the counting processes:

$$C_{q,i}\left(t\right) = \text{number of observed transitions of type } q \text{ for } i \text{ in } \left[0, t\right],$$

described by the times $t_{q,i,c}$ of these transitions, with:

$$0 < t_{q,i,1} < \ldots < t_{q,i,C_{q,i}(t_{\max,i})} {\leq} t_{\max,i}.$$

We denote the overall transition counting process by $C_q\left(t\right) = \sum\limits_{i=1}^{N} C_{q,i}\left(t\right)$. We will furthermore need an at-risk indicator process for transition $q = k.k'$ which we define by:

$$R_{q,i}\left(t\right) = \mathrm{I}_{\{Y_i(t-)=k\}},$$

and $R_q\left(t\right) = \sum\limits_{i=1}^{N} R_{q,i}\left(t\right)$, with $C_{q,i}\left(t\right) = C_{q,i}\left(t_{\max,i}\right)$ and $R_{q,i}\left(t\right) = 0$ for $t > t_{\max,i}$. The at-risk processes $R_q\left(t\right)$ and $R_{q,i}\left(t\right)$ are identical across $t \in \mathcal{T}$ for all $q \in \mathcal{Q}$ with the same entry state $k \in \mathcal{K}$.
The likelihood – abbreviated in the following by Lik –, conditional on the initial distribution of the multi-state process and the density of covariates, is [14]:

$$\mathrm{Lik} = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \left(\prod_{q=1}^{Q} \left[\exp\left(- \int\limits_{0}^{t_{\max,i}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \prod_{c=1}^{C_{q,i}(t_{\max,i})} \lambda_{q,i}\left(t_{q,i,c}\right) \right] \right).$$

We may rewrite each individual likelihood contribution Lik_i as:

$$\mathrm{Lik}_i = \prod_{q=1}^{Q} \left[\lambda_{q,i}\left(t_{q,i,1}\right) \exp\left(- \int\limits_{0}^{t_{q,i,1}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \cdot \right.$$

$$\left. \cdot \prod_{c=2}^{C_{q,i}(t_{\max,i})} \lambda_{q,i}\left(t_{q,i,c}\right) \exp\left(- \int\limits_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \right].$$

With this formulation, "two patterns of incomplete observations are particularly easy tractable" [4]: independent right-censoring and left-truncation (see Section 2.2.1 in Beyersmann *et al.* [15]). For left-truncation, only the lower integral boundary has to be changed from the value 0 to the time of delayed entry, which we denote by $t_{q,i,0}$:

$$\mathrm{Lik}_i = \prod_{q=1}^{Q} \left[\lambda_{q,i}\left(t_{q,i,1}\right) \exp\left(- \int\limits_{t_{q,i,0}}^{t_{q,i,1}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \cdot \right.$$

$$\cdot \prod_{c=2}^{C_{q,i}(t_{\max,i})} \lambda_{q,i}\left(t_{q,i,c}\right) \exp\left(-\int\limits_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t\right)\Bigg] .$$

For right censoring, a non-censoring indicator δ_i has to be included for the potential last jump of the counting processes at $t_{\max,i}$:

$$\mathrm{Lik}_i = \prod_{q=1}^{Q} \Bigg[\left(\prod_{c=1}^{C_{q,i}(t_{\max,i})-1} \lambda_{q,i}\left(t_{q,i,c}\right) \exp\left(-\int\limits_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t\right)\right) \cdot$$

$$\cdot \lambda_{q,i}\left(t_{\max,i}\right)^{\delta_i} \exp\left(-\int\limits_{t_{q,i,C_{q,i}\left(t_{\max,i}\right)-1}}^{t_{\max,i}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t\right)\Bigg] .$$

So we can treat this as a combined product over c where a transition-specific non-censoring indicator $\delta_{q,i,c}$ is always equal to one despite $\delta_{q,i,C_{q,i}(t_{\max,i})}$ which may also be equal to zero:

$$\mathrm{Lik}_i = \prod_{q=1}^{Q} \left[\prod_{c=1}^{C_{q,i}(t_{\max,i})} \left(\lambda_{q,i}\left(t_{q,i,c}\right)^{\delta_{q,i,c}} \exp\left(-\int\limits_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t\right)\right)\right] . \qquad (1)$$

2.2. Parametrization of transition-specific hazard rate functions

In event-time analysis, statistical models are often obtained by specifying transition-specific hazard rate functions $\lambda_{q,i}\left(t\right)$ for each individual i. The most widely used models have a multiplicative structure with a transition-specific baseline hazard rate function $\lambda_{q,0}\left(t\right)$. For an individual i, the transition-specific baseline hazard rate is then modelled by [1, 14]:

$$\lambda_{q,i}\left(t\right) = \lambda_{q,0}\left(t\right) \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}_q\right),$$

with time-constant covariates $\mathbf{x}_i = \left(x_{1,i}, \ldots, x_{P,i}\right)^\top$ and respective effects $\boldsymbol{\beta}_q = \left(\beta_{q,1}, \ldots, \beta_{q,P}\right)^\top$ on transition q. Their product is the transition-specific linear predictor $\eta_{q,i} = \mathbf{x}_i^\top \boldsymbol{\beta}_q$. Hence, the effect of a covariate x_p is described by factors $\exp\left(\beta_{q.p}\right)$ that proportionally modify the transition-specific baseline-hazard rate function $\lambda_{q,0}\left(t\right)$.

For the commonly applied continuous time Markov model, the multi-state process $Y\left(t\right)$ is a Markov process, i.e. takes the assumption that the dependence of the transition-specific hazard rate functions $\lambda_q\left(t\right)$ on the history $\mathcal{Y}_t$ is only via the current state of $Y\left(t\right)$ and possibly via time-fixed covariates. For notational simplicity, we only use time-constant covariates here (see Cortese and Andersen [16] for a detailed description of how to include time-dependent covariate information such as duration lengths).

In the simplest version, the transition-specific baseline hazard rates are kept constant:

$$\lambda_{q,0}\left(t\right) = \lambda_{q,0},$$

or piecewise constant:

$$\lambda_{q,0}\left(t\right) = \lambda_{q,0}^{(j)}, \quad t^{(j-1)} < t \leq t^{(j)},$$

for a time-axis decomposed into several sub-intervals by artificial time points $t^{(0)} = 0, t^{(1)}, t^{(2)}, \ldots, t^{(j)}, \ldots t^{(J)} = t_{\max}$.
The Cox partial likelihood model [1] leaves the baseline hazard rate functions unspecified, but assumes them to be equal across individuals. If one is not interested in the underlying functional form of baseline hazard rate function, the Cox Partial likelihood model is a good choice since it leaves no room for functional mis-specification.
We will use both, a piecewise constant, and an unspecified hazard rate function parametrisation, to set up fusion Lasso penalised multi-state models. Details on a stratified Cox Partial likelihood formulation for multi-state models are given in Appendix B, details to the piecewise constant set-up – which is referred to in the literature as *Piecewise Exponential Model* – in Appendix E. Of course, many parametric alternatives of the baseline hazard rate function specification exist, but will not be treated in this article. In general, all of these models need a pre-estimation data management procedure that will be described in Appendix A.

2.3. Penalised likelihood formulation for multi-state models and the fusion Lasso penalty

A general penalised negative log (Partial) likelihood is defined by:

$$\text{PenNegLog(Partial)Lik}\left(\boldsymbol{\beta}\right) = -\text{Log(Partial)Lik}\left(\boldsymbol{\beta}\right) + \text{pen}\left(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}\right).$$

Here, the penalty $\text{pen}\left(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}\right)$ may be defined in different ways, depending on which features of the covariate mechanism shall be detected. Technically, different penalties are achieved by a penalty structure matrix $\mathbf{D}$ which is described in the following, accompanied by the introduction of different types of penalty structures.
The *least absolute shrinkage and selection operator*, short *Lasso*, introduced by Tibshirani [7], maximizes a likelihood "subject to the sum of the absolute value of the coefficients being less than a constant" [7]. A Lasso type penalty term penalising the absolute value of all elements of the parameter vector $\boldsymbol{\beta}$ is therefore constructed as:

$$\text{pen}_{\text{L}}\left(\lambda, \mathbf{D}, \boldsymbol{\beta}\right) = \lambda \sum_{p=1}^{P} |\beta_p| = \lambda \sum_{l=1}^{P} |\mathbf{d}_l^{\top} \boldsymbol{\beta}|,$$

using the penalty parameter λ, parameter vector $\boldsymbol{\beta}$, and difference vectors $\mathbf{d}_l^{\top} = (0, \ldots, 0, 1, 0, \ldots, 0)$ taking value 1 in the l-th entry, and 0 otherwise. The vectors $\mathbf{d}_l$ are then stored as lines in the penalty structure matrix $\mathbf{D}$, which is here equal to the $P \times P$ dimensional identity matrix. The matrix $\mathbf{D}$ is useful to incorporate several penalty types into one unifying approach. This will become clearer in the following paragraph.

The *Fused Lasso* [17] was introduced with the intention to generalize the Lasso penalisation approach "for problems with features that can be ordered in some meaningful way" [17]. This is achieved by penalising the L_1-norm, i.e. the absolute value, of both the coefficients and their successive differences. By this, the Fused Lasso leads to sparsity of the coefficients and also of the differences between adjacent covariate level effects. A Fused Lasso type penalty term penalising the absolute value of all elements of the parameter vector $\boldsymbol{\beta}$ and all of the successive differences is constructed as:

$$\mathrm{pen}_{\mathrm{FL}}\left(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}\right) = \mathrm{pen}_{\mathrm{L}}\left(\lambda_1, \mathbf{D}_{\mathrm{L}}, \boldsymbol{\beta}\right) + \mathrm{pen}_{\mathrm{F}}\left(\lambda_2, \mathbf{D}_{\mathrm{F}}, \boldsymbol{\beta}\right),$$

with $\mathrm{pen}_{\mathrm{F}}\left(\lambda_2, \mathbf{D}_{\mathrm{F}}, \boldsymbol{\beta}\right) = \lambda_2 \sum\limits_{p=1}^{P-1} |\beta_{p+1} - \beta_p| = \lambda_2 \sum\limits_{l=1}^{P-1} |\mathbf{d}_{\mathrm{F},l}^{\top}\boldsymbol{\beta}|$, fusion difference vectors $\mathbf{d}_{\mathrm{F},l}^{\top} = (0, \ldots, 0, -1, 1, 0, \ldots, 0)$, penalty parameter vector $\boldsymbol{\beta}$, and a combined penalty structure matrix $\mathbf{D} = \left[\mathbf{D}_{\mathrm{L}}, \mathbf{D}_{\mathrm{F}}\right]^{\top}$, with the fusion difference vectors $\mathbf{d}_{\mathrm{F},l}^{\top}$ stored as lines in the penalty structure matrix $\mathbf{D}_{\mathrm{F}}$.

The *Pairwise Fused Lasso* proposed by Petry *et al.* [18] extends the Fused Lasso [17] to models where the predictors have no natural ordering. Here not only next neighbour coefficient differences, but all pairwise coefficient differences are penalised. As a consequence, the fusion difference vectors $\mathbf{d}_{\mathrm{F},l}^{\top} = (0, \ldots, 0, -1, 1, 0, \ldots, 0)$ for differences between adjacent effects are supplemented by the remaining vectors $\mathbf{d}_{\text{all pairwise fusion},l}$ leading to all possible pairwise effect differences.

To penalise transition-specific covariate coefficients and their pairwise differences in multi-state modelling, a Fusion Lasso penalty term will be introduced. This penalty term takes into account the information provided by the state-chart of the multi-state model and is of the general form:

$$\mathrm{pen}_{\mathrm{SFL}}\left(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}\right) = \lambda_1 \sum_{q=1}^{Q} \sum_{p=1}^{P} |\beta_{p,q}| + \lambda_2 \sum_{q,q'} \sum_{p=1}^{P} |\beta_{p,q} - \beta_{p,q'}|,$$

with $\lambda_1 \sum\limits_{q=1}^{Q} \sum\limits_{p=1}^{P} |\beta_{p,q}|$ as Lasso type and $\lambda_2 \sum\limits_{q,q'} \sum\limits_{p=1}^{P} |\beta_{p,q} - \beta_{p,q'}|$ as fusion type penalty term. Here, the indices q, q' denote suitable pairs of transitions that are to be extracted from the state-chart of the multi-state model.

One important aspect is that the solutions of penalisation approaches, as e.g. Lasso, are not equivariant under scaling of the covariates [19]. In other words, there is a dependency of the solution of the penalised estimation approach with respect to the scales of the covariates when unique penalty parameter values λ_1 and λ_2 are selected for both penalty term components. A frequently used solution to get rid of this general problem in penalisation algorithms is to use standardized covariate versions [19], i.e. $\mathbf{x}_p^* := \frac{\mathbf{x}_p - \hat{\mu}_{\mathbf{x}_p}}{\hat{\sigma}_{\mathbf{x}_p}}$, with $\hat{\mu}_{\mathbf{x}_p}$ as the empirical mean of $\mathbf{x}_p$, and $\hat{\sigma}_{\mathbf{x}_p}$ as the empirical standard deviation of $\mathbf{x}_p$. For the interpretation, the coefficients are back-transformed after the estimation is performed. It is important to note here that this scaling has to be performed on combined transitions $q.q'$, i.e. $\hat{\mu}_{\mathbf{x}_{p,q.q'}}$ and $\hat{\sigma}_{\mathbf{x}_{p,q.q'}}$, to maintain the fusion feature for coefficient estimates of covariate x_p and transitions q and q' after the back-transformation step. This has the consequence that not all covariates within the penalised estimation task are having exactly the same scaling. However, there is no escape from this problem: scaling each covariate independently , i.e. $\mathbf{x}^*_{p,q} := \frac{\mathbf{x}_{p,q} - \hat{\mu}_{\mathbf{x}_{p,q}}}{\hat{\sigma}_{\mathbf{x}_{p,q}}}$ and $\mathbf{x}^*_{p,q'} := \frac{\mathbf{x}_{p,q'} - \hat{\mu}_{\mathbf{x}_{p,q'}}}{\hat{\sigma}_{\mathbf{x}_{p,q'}}}$, leads to equal scales, group scaling leads to the preservation of the fusion feature. For the peritoneal dialysis program data that will be analysed in Section 3, the range for standard deviations of the scaled covariates $\mathbf{x}^*_p$ ranges between 0.87 and 1.1, and we therefore consider these differences to be of only low relevance. This is even more the case given the fact that the frequencies of transition observation are quite unbalanced in this example. A heuristic way of controlling whether these results lead to any false effect fusions is to re-estimate the model with individually scaled covariates and check if any different effect fusions occur.

2.4. Basic estimation algorithms and the penalised iterative re-weighted least squares algorithm

The approach introduced by this article may be performed on the log partial likelihood $\text{LogPartialLik}\left(\boldsymbol{\beta}\right)$ or the log likelihood $\text{LogLik}\left(\boldsymbol{\beta}\right)$, where the conditioning on other quantities is still notationally suppressed as described in Appendix B and Appendix E.

A naive way of getting penalised estimates would be to directly minimize the penalised negative log (partial) likelihood $\text{PenNegLog(Partial)Lik}\left(\boldsymbol{\beta}\right)$ with respect to the parameters $\boldsymbol{\beta}$, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\text{PenNegLog(Partial)Lik}\left(\boldsymbol{\beta}\right)\right),$$

using direct numerical optimization techniques, e.g. the Nelder-Mead [20] algorithm. Since the performance of this would not be optimal with respect to computational cost and instability, we rather rely on a modified version of an algorithm [8] which is based on a first order Taylor series expansion $\left(\mathbf{z} - \boldsymbol{\eta}\right)^\top \mathbf{A}\left(\mathbf{z} - \boldsymbol{\eta}\right)$ for the log (partial) likelihood $\text{Log(Partial)Lik}\left(\boldsymbol{\beta}\right)$, with $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{u} = \frac{\partial\, \text{Log(Partial)Lik}}{\partial\, \boldsymbol{\eta}}$, $\mathbf{A} = -\frac{\partial^2\, \text{Log(Partial)Lik}}{\partial\, \boldsymbol{\eta}\boldsymbol{\eta}^\top}$, and $\mathbf{z} = \boldsymbol{\eta} - \mathbf{A}^{-1}\mathbf{u}$. The first and second derivatives of the $\text{LogPartialLik}\left(\boldsymbol{\beta}\right)$ with respect to the linear predictor $\boldsymbol{\eta}$ are given e.g. in Hastie and Tibshirani [21]. If the calculation of $\mathbf{A}$ is computationally very burdensome, as it is in the partial likelihood case, one is able to use a reduced version of $\mathbf{A}$ that contains the same diagonal elements $a_{i,i}$, but is equal to 0 in all other entries [21].

An iterative minimization is then performed by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \Big(\left(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\right)^\top \mathbf{A}\left(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\right) + \text{pen}\left(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}\right)\Big),$$

where $\mathbf{z}$ and $\mathbf{A}$ are calculated using the version of $\boldsymbol{\beta}$ from the respective previous iteration. Before the first iteration, we set $\hat{\boldsymbol{\beta}} = (0, \ldots, 0)^\top$. This algorithm is iteratively performed using numerical optimization techniques for the nested minimization in each step. The algorithm is pursued until an updated version of $\hat{\boldsymbol{\beta}}$ – for the first time – does not change any more when compared to its previous version (with respect to a certain tolerance, e.g. $\frac{\sum_j |\hat{\beta}_{\text{old}} - \hat{\beta}_{\text{new}}|}{\sum |\hat{\beta}_{\text{new}}|} < 10^{-5}$).

An alternative, generally known, easy to implement, and therefore preferable, estimation approach is the penalised iteratively re-weighted least squares (PIRLS) algorithm [22]. The iteratively re-weighted least squares (IRLS) algorithm is very familiar since it is used in generalized linear models to find the maximum likelihood estimates. The PIRLS approach provides an estimation framework that is build-up on the well established IRLS basis, gives flexibility to incorporate penalties and yields very stable results – equal to those of suitably specified reference software implementations [22].

To use the PIRLS algorithm for multi-state models, it is required to calculate the score vector and Fisher information as described in later parts of this section. Furthermore we need a local quadratic approximation $\mathbf{P}_{\boldsymbol{\lambda}}$ of the penalty matrix generally defined as [22]:

$$\mathbf{P}_{\boldsymbol{\lambda}} = \sum_{l=1}^{L} \lambda_l \mathbf{P}_l = \sum_{l=1}^{L} \lambda_l \frac{\partial \xi \left(||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l} \right)}{\partial \, ||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}} \frac{\mathcal{D}_l \left(\mathbf{d}_l^\top \boldsymbol{\beta} \right)}{\mathbf{d}_l^\top \boldsymbol{\beta}} \mathbf{d}_l \mathbf{d}_l^\top .$$

Here, a penalty function ξ of the form $\xi \left(||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l} \right) = ||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}$ is used. Consequently the derivative neutralises, since $\frac{\partial \xi \left(||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l} \right)}{\partial \, ||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}} = 1$. Without exception, we rely on penalty terms that penalise the L_1-norm, also known as *absolute value function*, i.e. $||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l} = ||\mathbf{d}_l^\top \boldsymbol{\beta}||_1 = |\mathbf{d}_l^\top \boldsymbol{\beta}|$. A quadratic approximation $\mathcal{N}_{\text{L}_1}$ to this L_1 norm is $\mathcal{N}_{\text{L}_1} \left(\mathbf{d}_l^\top \boldsymbol{\beta} \right) = \sqrt{\left(\mathbf{d}_l^\top \boldsymbol{\beta} \right)^2 + c}$ [22], with derivative $\mathcal{D}_{\text{L}_1} \left(\mathbf{d}_l^\top \boldsymbol{\beta} \right) = \frac{\mathbf{d}_l^\top \boldsymbol{\beta}}{\sqrt{\left(\mathbf{d}_l^\top \boldsymbol{\beta} \right)^2 + c}}$, where c is a very small constant (we use $c = 10^{-8}$ as suggested by Petry *et al.* [18]). This leads to the following special form of the quadratic approximation $\mathbf{P}_{\boldsymbol{\lambda}}$ of the penalty matrix

$$\mathbf{P}_{\boldsymbol{\lambda}} = \sum_{l=1}^{L} \lambda_l \mathbf{P}_l = \sum_{l=1}^{L} \lambda_l \cdot \frac{\frac{\mathbf{d}_l^\top \boldsymbol{\beta}}{\sqrt{\left(\mathbf{d}_l^\top \boldsymbol{\beta} \right)^2 + c}}}{\mathbf{d}_l^\top \boldsymbol{\beta}} \mathbf{d}_l \mathbf{d}_l^\top .$$

The construction of the penalty structure vector $\mathbf{d}_l$ is of the type $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$ to penalise a single effect (Lasso term) and of the type $(0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^\top$ to penalise the difference between two effects (fusion term), with Lasso terms and fusion terms as described in Section 2.3.

For the estimation of penalised models, we need the score vector $\mathbf{s}(\boldsymbol{\beta}) = \frac{\partial \, \text{LogPartialLik}(\boldsymbol{\beta})}{\partial \, \boldsymbol{\beta}}$ of the log Partial likelihood with components $s_p(\boldsymbol{\beta}) = \frac{\partial \, \text{LogPartialLik}(\boldsymbol{\beta})}{\partial \, \beta_p}$:

$$s_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\delta_i x_{i,p} - \delta_i \sum_{j \in R_i} \frac{\exp(\eta_j) \, x_{j,p}}{\sum\limits_{k \in R_i} \exp(\eta_k)} \right),$$

and the Fisher information matrix $\mathbf{F}(\boldsymbol{\beta}) = \dfrac{\partial^2 \, \text{LogPartialLik}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^\top}$ with components $F_{p,p'}(\boldsymbol{\beta}) = \dfrac{\partial^2 \, \text{LogPartialLik}(\boldsymbol{\beta})}{\partial \beta_p \, \partial \beta_{p'}}$:

$$F_{p,p'}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \frac{\sum\limits_{j \in R_i} \exp(\eta_j) \, x_{j,p} x_{j,p'}}{\sum\limits_{k \in R_i} \exp(\eta_k)} - \sum_{i=1}^{n} \delta_i \frac{\left(\sum\limits_{j \in R_i} \exp(\eta_j) \, x_{j,p} \right) \cdot \left(\sum\limits_{j \in R_i} \exp(\eta_j) \, x_{j,p'} \right)}{\left(\sum\limits_{k \in R_i} \exp(\eta_k) \right)^2} .$$

The PIRLS algorithm is then composed in the following way (using a step-length factor $\nu \in (0,1]$ and iteration counter h) [22]:

$$\hat{\boldsymbol{\beta}}_{(h+1)} = \hat{\boldsymbol{\beta}}_{(h)} - \nu \cdot \left(-\mathbf{F}\left(\hat{\boldsymbol{\beta}}_{(h)} \right) - \mathbf{P}_{\boldsymbol{\lambda}} \right)^{-1} \left(\mathbf{s}\left(\hat{\boldsymbol{\beta}}_{(h)} \right) - \mathbf{P}_{\boldsymbol{\lambda}} \hat{\boldsymbol{\beta}}_{(h)} \right) .$$

Again, this algorithm is terminated when the relative successive differences between the estimated coefficients is smaller than a fixed convergence criterion [22]. We define the starting vector as $\boldsymbol{\beta} = (0, \ldots, 0)^\top$. Using several different starting values is a good way to prevent the algorithm from running into local optima, a problem that has never occurred during the research process leading to this article.

2.5. Selection of penalty parameters

Several alternative criteria for tuning parameter or model selection exist across the literature. One of the most frequently used criteria to select optimal penalty parameters $\boldsymbol{\lambda}$ is the *Akaike Information Criterion (AIC)* defined by:

$$\text{AIC} = -2 \cdot \text{Log(Partial)Lik} + 2 \cdot \text{df}.$$

For the calculation of the AIC, we require a measure for the model complexity, i.e. the model degrees of freedom (df). In analogy to an article on the estimation of non-linear covariate effects in a Cox Proportional Hazards model using penalised splines [23], the model degrees of freedom are calculated by:

$$\text{df} = \text{trace}\left((\mathbf{F} + \mathbf{P}) (\mathbf{F} + \mathbf{P})^{-1} \mathbf{F} (\mathbf{F} + \mathbf{P})^{-1} \right) = \text{trace}\left(\mathbf{F} (\mathbf{F} + \mathbf{P})^{-1} \right) ,$$

where $\mathbf{F}$ is the Fisher information, and $\mathbf{P}$ is the second derivative matrix of the penalty function.

An alternative definition is established by Tibshirani *et al.* [17] in the fused Lasso context with:

$$\text{df} = p - \# \{ \beta_j = 0 \} - \# \{ \beta_j - \beta_{j-1} = 0; \beta_j, \beta_{j-1} \neq 0 \} .$$

In other words, this definition is to “count a sequence of one or more consecutive non-zero and equal β_j-values as one degree of freedom” [17]. Gertheiss and Tutz [24] slightly

adapt this definition to ordinal penalties by counting the number of unique non-zero coefficients induced by the respective current parameter estimates. However, we prefer to use the approach using the Fisher information matrix, since it has the advantage that the degrees of freedom formulation is a continuous function of $\boldsymbol{\lambda}$, while the definitions by Tibshirani *et al.* [17] and Gertheiss and Tutz [24] introduce discontinuous step functions. Moreover it corresponds very naturally to the PIRLS approach introduced in Section 2.4.
Using a grid search is a frequently applied strategy for the selection of a pair of tuning parameters, see for example Tibshirani *et al.* [17] or Zou and Hastie [25]. This is pursued by taking all pair-wise combinations constructed by candidate vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ and then choosing the best combination (λ_1, λ_2) amongst all pair-wise combinations of tuning parameter values with respect to a selection criterion. Typically one would set-up a candidate grid using a finer lower part and then checking on a coarser grid for larger values, e.g. by using candidate vectors $\boldsymbol{\lambda}_l = (0, 0.01, 0.1, 1, 2, 3, \ldots, 10, 50, 100)$, $l = 1, 2$.

2.6. Software implementation

The established methods are implemented in the R [10] add-on package `penMSM` [26]. The central function within this package is named `penMSM` and performs the PIRLS algorithm established in Section 2.4. Besides other mandatory objects, it is key to forward the penalty structure matrices `PSM1` (Lasso part) and `PSM2` (fusion part) as well as the vectors with the penalty parameters for the respective penalty components (`lambda1` for the Lasso part and `lambda2` for the fusion part) to the `penMSM` function. Examples for these object definitions will be given in Appendix D.

3. Structured fusion Lasso estimation of a four state model for peritoneal dialysis program data

The potential of structured fusion Lasso estimation for multi-state models will be demonstrated during this section by analysing peritoneal dialysis study data (monthly time scale) for chronic renal disease patients as an application example. The goal of this section is not to perform an analysis that is completely adequate from a strictly medical point of view but rather to illustrate the potential of the structured fusion Lasso to obtain sparse models in the multi-state context.
Peritoneal dialysis is a class of dialysis methods that has important advantages in comparison to other dialysis methods like Haemodialysis. Some of these advantages are a longer salvage of the remaining renal function, less frequent complications with respect to the dialysis access, and greater independence of patients from dialysis centres – eligible patients can independently carry out the treatment at home which results in a boost of life quality, for example patients are still able to travel. However, a major disadvantage of the peritoneal dialysis is a higher risk that the abdominal cavity is infected with pathogenic bacteria when getting into contact with the environment, with peritonitis – an inflammation of the peritoneum, the thin tissue that lines the inner wall of the abdomen and covers most of the abdominal organs – as the possible consequence.

Therefore, patients must work very carefully and as sterile as possible when changing the dialysis solutions. Moreover, since peritoneal dialysis uses sugar-based solutions to perform dialysis, patients affected by diabetes will have to additionally adapt their diabetic medication.

In general, end stage renal disease is a worldwide increasing health problem, with a considerable amount of patients in need of a renal replacement treatment or having some degree of renal dysfunction [27]. Moreover, the complications of "diabetes and hypertension are the two most common causes of end stage renal disease and are associated with a higher risk of death from cardiovascular disease" [27]. Hence, increasing the knowledge about the underlying mechanisms that lead to different complications and complication sequences after starting a peritoneal dialysis program is of high interest. The states in the peritoneal dialysis program study have already been described in Section 1.

Figure 2 shows the state-chart that illustrates the possible seven transitions between the four states, where this four state process splits up into two nested competing risk models, both containing transitions to the absorbing states D, H, and R. The numbers of observed transitions with entrance to the study (E) as initial state are 215for the transition to peritonitis (EP), 47 transitions to death (ED), 56 transitions to transfer to haemodialysis (EH), 67 transitions to renal transplantation (ER), and 40 right-censored observations. For the transitions with peritonitis (P) as initial state, 47 transitions to death (PD), 94 transitions to transfer to haemodialysis (PH), and 48 transitions to renal transplantation (PR) have been observed, with 26 right-censored observations. Figure 3 gives Nelson-Aalen estimates for the seven transitions in the peritoneal dialysis program data. The transition-specific median sojourn times given in months since entry to the study are illustrated as vertical lines in the top of Figure 8.

Four personal or clinical characteristics are taken into account for possibly influencing the transition-specific hazard rate functions:

- Age of the patient at entry into the study (*Age*, measured in years),
- Sex of the patient (*Sex*, male/female, with reference category defined as female),
- *Diabetes* (no diabetes as reference category), and
- Previous renal transplantation therapy (*PRRT*, no PRRT as reference category).

The age of the patients has been taken into account with a potentially non-linear effect based on a fractional polynomial as described for the baseline hazard rate specification in Appendix E. Appendix D gives some code snippets on how to set up design and penalty matrices for the piecewise exponential modelling in the peritoneal dialysis application example.

This section presents the results for the structured fusion Lasso penalized Cox model approach, while Appendix E gives the results for the piecewise exponential model which additionally specifies the transition-specific baseline hazard rate functions. Both approaches are compared to an un-penalised multi-state model which is estimated using the benchmark software `BayesX` [29, 30]. Here, transition-specific effects and 95%

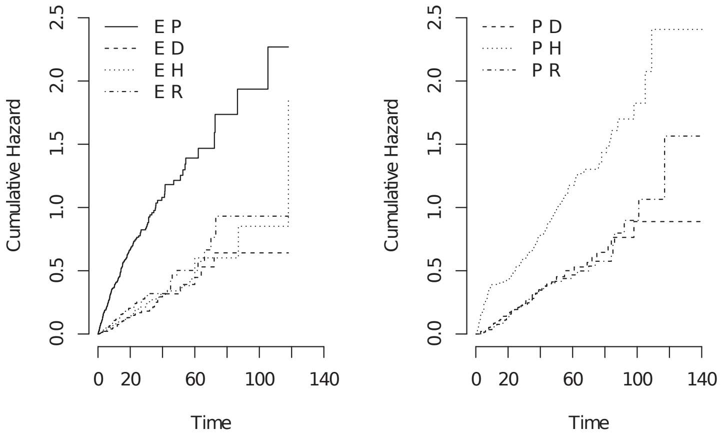

Figure 3: Nelson-Aalen estimates for the transitions in the peritoneal dialysis program data (time in months). Estimates have been calculated using R add-on package `mvna` [28]. The left panel shows estimates for the transitions from entrance (E) to the study, the right panel shows estimates for the transitions from the affection with peritonitis (P).

confidence intervals are determined for the baseline hazard rate functions, effects of age, sex, diabetes and PRRT. Furthermore, the non-linear functional form of the baseline hazard rate functions and the effects of age are estimated using penalised B-Splines [31]. This model has been similarly set up – using additional time-varying effects – by Teixeira *et al.* A direct unpenalized estimation, e.g. by the use of the R-function `coxph` from the `survival` package [33] using fractional polynomials for the specification of the age leads to convergence problems and is therefore not available for comparison purposes.

	E to P	E to D	E to H	E to R	P to D	P to H	P to R
Sex	-0.27 (0.15)	-0.36 (0.32)	0.26 (0.29)	0.17 (0.26)	0.17 (0.31)	-0.19 (0.23)	-0.58 (0.34)
Diabetes	-0.06 (0.18)	0.88 (0.32)	0.55 (0.32)	0.00 (0.33)	0.88 (0.33)	0.09 (0.27)	-0.20 (0.45)
PRRT	0.29 (0.14)	0.66 (0.31)	0.48 (0.29)	-0.27 (0.26)	0.58 (0.31)	0.11 (0.21)	-0.29 (0.30)

Table 1: Results for sex, diabetes and PRRT by Teixeira *et al.*: estimated transition-specific coefficients, with respective standard error in brackets.

Table 1 shows the estimated coefficients for the binary covariates sex, diabetes and PRRT by Teixeira *et al.* The results suggest to take into account possible sparsity on the individual level, and moreover fused pairs of effects, e.g. for the effects of diabetes on the transitions ED and PD ($\hat{\beta}_{\text{diab.},ED} = 0.88$, and $\hat{\beta}_{\text{diab.},PD} = 0.88$). Note here that the effect for diabetes on the transition ER is already equal to 0, but both of these parsimonious findings are results of rounding to the 2^{rd} digit since the approach is not

able to perform automated variable selection.
We estimate the structured fusion Lasso penalized Cox model on a grid of (λ_1, λ_2) combinations constructed by all pair-wise combinations for candidate vectors $\lambda_1 = \{0.1, 0.305, 0.511, \dots, 4\}$, and $\lambda_2 = \{0.1, 0.621, 1.142, \dots, 10\}$, as being generated as equidistant sequences. The resulting AIC values are illustrated in Figure 4. The minimal AIC with value 5335.456 is reached for the combination $(\lambda_1, \lambda_2) = (2.153, 7.395)$. Any value of (λ_1, λ_2) outside the illustrated region (results not shown) leads to a considerable increase of the AIC, and the minimum at $(2.153, 7.395)$ therefore stands for the global AIC minimum.
The results for the minimum AIC model are illustrated in Figure 5. We see that the effects for covariate sex are all equal to or smaller than 0. This yields a clearer image of the sex effect in comparison to the un-penalised estimation, where the effect was smaller than 0 for four transitions, and larger than 0 for three other transitions (denoted by the middle of the 95% confidence intervals). To be more precise, all of the effects larger than 0 in the un-penalised estimation are shrunken to 0, whereas all the effects smaller than 0 stay below and different to the value 0. Cross-transition effects are obtained for the transitions to death (for the effect of diabetes and PRRT) and for the transfer to renal transplantation (again for the effects of diabetes and PRRT). Note here that any small coefficient differences preventing setting an effect to 0 or fusing two effects appear also for penalty parameter combinations next to the minimum AIC combination $(\lambda_1, \lambda_2) = (2.153, 7.395)$. Hence they are not attributable to a too coarse grid of penalty parameter combination candidates. For the estimated age effects illustrated at the top of Figure 5, we find a positive, almost linear effect of age on the transition from entrance to peritonitis. We furthermore identify cross-transition age effects, i.e. a fusion of transition-specific age effects, for the transition combinations EH with PH, and ER with PR. The difference between the estimated age effects for the transition combination ED with PD is reduced to almost 0, but the effects are not yet fused at $(\lambda_1, \lambda_2) = (2.153, 7.395)$. We therefore get strong evidence that age is not associated with different hazards of transition to the competing endpoints in the study whether a first peritonitis occurred or not.

4. Discussion

With the structured fusion Lasso penalised multi-state modelling approach introduced by this article, we establish a data-driven way to perform a structured analysis of multi-state models with potential cross-transition effect and variable selection. We presented the approach for partial likelihood and piecewise constant baseline hazard rate models and proposed an algorithm that is applicable to a broad class of multi-state models. This is achieved by the use of a penalised iterative reweighed least squares algorithm that is close to estimation algorithms known from generalized linear models. We are able to use this algorithm in combination with a local quadratic approximation of the L_1-norm. The best combination of Lasso and fusion penalty parameters is selected using a grid search for the minimal Akaike information criterion value.

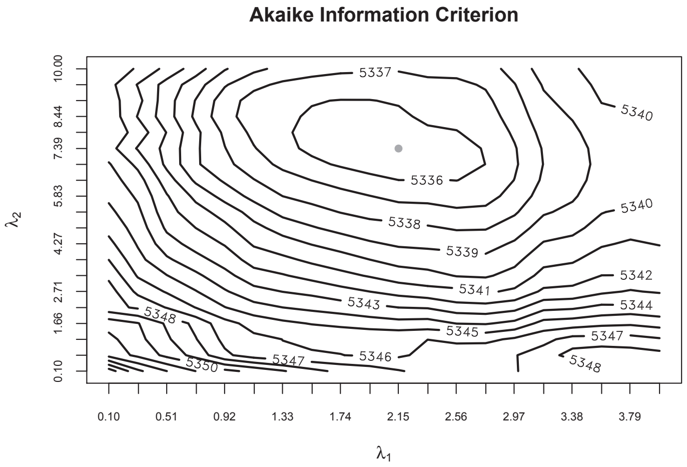

Figure 4: Contour plot of the Akaike Information Criterion values on the (20×20)-grid constructing combinations of the penalty parameters (λ_1, λ_2). The black lines denote contour lines of the two dimensional AIC surface, the grey point indicates the $(\lambda_1, \lambda_2) = (2.153, 7.395)$ coordinate with the minimal AIC of 5335.46.

The application to peritoneal dialysis data showed that we are able to work out interesting insights into the structural relationships between transitions, a problem that has been addressed several times in multi-state modelling literature, but has never been entrusted with a suitable data-driven estimation concept.

One potential direction of future research is to generalise the fusion of effects to nonlinear effects represented, for example, as penalised B-splines. However, this will considerably increase the number of penalty parameters involved such that more automatic ways of estimating these jointly with the regression coefficients are desired. This could for example be accomplished in a Bayesian treatment of the structured fusion Lasso where suitable priors are assigned to the penalty parameters.

Acknowledgement: We wish to thank Laetitia Teixeira, Anabela Rodrigues, and Denisa Mendonça from the University of Porto, Portugal, and Carmen Cadarso-Suárez from the University of Santiago de Compostela, Spain, for kindly providing us the data used for the peritoneal dialysis program data application. The authors have been supported by the German Research Foundation (DFG) research training group 1644 Scaling Problems in Statistics.

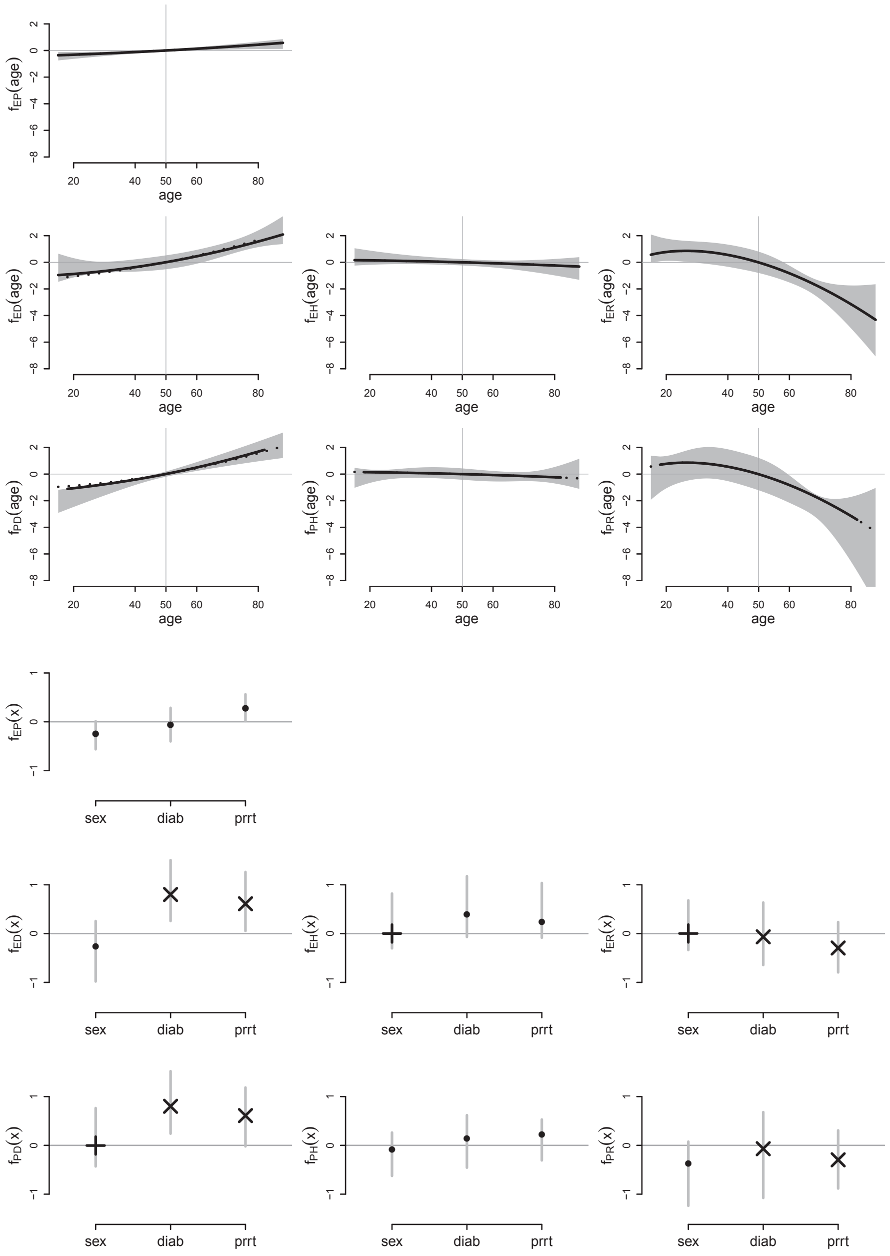

Figure 5: Estimated transition-specific effects of age (seven plotting windows at the top), sex, diabetes, and PRRT (seven plotting windows at the bottom) for the structured fusion Lasso penalised Cox model in the peritoneal dialysis program data application with the minimum AIC penalty parameter combination $(\lambda_1, \lambda_2) = (2.153, 7.395)$.

Age effect illustrations: Since we penalize the difference between effects of the same covariate on pairs of transitions, the fusing strength is made visible by combining pairs of effects in the respective plotting windows. Solid black lines illustrate the respective estimated effect for the transitions as annotated in the axis labels, dotted illustrate denote the estimate of the respective associated effect function. Grey areas show point-wise 95% confidence intervals of the benchmark model using the software `BayesX`, where smooth effect estimates are obtained using penalised B-Splines. The estimated effects and confidence intervals are centred around an age of 50 years.

Sex, diabetes, and PRRT effect illustrations: black plotting symbols denote the estimated effect values, grey vertical lines illustrate 95% confidence intervals of the `BayesX` benchmark model. Black bullet points (•) denote non-fused effects different to 0, plus signs (+) denote effects equal to 0, crosses (×) denote fused effects. Combinations of a plus sign and a cross appear as stars

References

1. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) 1972; **34**(2):187–220.
2. Thall PF, Lachin JM. Assessment of stratum-covariate interactions in cox's proportional hazards regression model. Statistics in Medicine 1986; **5**(1):73–83, DOI: 10.1002/sim.4780050110.
3. Carstensen B, Plummer M. Using Lexis Objects for Multi-State Models in R. Journal of Statistical Software 2011; **38**(6):1–18.
4. Andersen PK, Keiding N. Multi-state models for event history analysis. Statistical Methods in Medical Research 2002; **11**(2):91–115, DOI:10.1191/0962280202SM276ra.
5. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part IV: Further concepts and methods in survival analysis. British Journal of Cancer 2003; **89**:781–786, DOI:10.1038/sj.bjc.6601117.
6. Schmidtmann I, Elsäß er A, Weinmann A, Binder H. Coupled variable selection for regression modeling of complex treatment patterns in a clinical cancer registry. Statistics in Medicine 2014; **33**(30):5358–5370, DOI:10.1002/sim.6340.
7. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996; **58**(1):267–288, DOI:10.1111/j.1467-9868.2011.00771.x.
8. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine 1997; **16**(4):385–395, DOI:10.1002/(SICI)1097-0258(19970228)16:4⟨385::AID-SIM380⟩3.0.CO;2-3.
9. Goeman JJ. L1 Penalized Estimation in the Cox Proportional Hazards Model. Biometrical Journal 2010; **52**(1):70–84, DOI:10.1002/bimj.200900028.
10. R Development Core Team. R: A Language and Environment for Statistical Computing. Software published online on the Comprehensive R Archive Net- work 2014; .
11. Goeman JJ. Penalized R package. The Comprehensive R Archive Network 2012; **version 0.9-42**.
12. Huang J, Zhang T. The benefit of group sparsity. The Annals of Statistics 8 2010; **38**(4):1978–2004, DOI:10.1214/09-AOS778.
13. Puig AT, Wiesel A, Fleury G, Hero AO. Multidimensional Shrinkage-Thresholding Operator and Group LASSO Penalties. Signal Processing Letters, IEEE 6 2011; **18**(6):363–366, DOI:10.1109/SSP.2009.5278625.
14. Andersen PK, Borgan O, Gill RD, Keiding N. Statistical Models Based on Counting Processes. Springer Series in Statistics, 1993.
15. Beyersmann J, Schumacher M, Allignol A. Competing Risks and Multistate Models with R. Springer Series "UseR!", 2012.
16. Cortese G, Andersen PK. Competing risks and time-dependent covariates. Biometrical Journal 2010; **52**(1), DOI:10.1002/bimj.200900076.
17. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness

via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2005; **67**(1):91–108, DOI:10.1111/j.1467-9868.2005.00490.x.

18. Petry S, Flexeder C, Tutz G. Pairwise Fused Lasso. Department of Statistics, University of Munich, Germany: Technical Reports 2011; **102**.
19. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., 2001.
20. Nelder JA, Mead R. A simplex method for function minimization. Computer journal 1965; **7**(4):308–313, DOI:10.1093/comjnl/7.4.308.
21. Hastie TJ, Tibshirani RJ. Generalized additive models, vol. 43. CRC Press, 1990.
22. Oelker MR, Tutz G. A uniform framework for the combination of penalties in generalized structured models. Advances in Data Analysis and Classification 2015; DOI:10.1007/s11634-015-0205-y.
23. Gray RJ. Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. Journal of the American Statistical Association 1992; **87**(420):942–951, DOI:10.1080/01621459.1992.10476248.
24. Gertheiss J, Tutz G. Sparse modeling of categorial explanatory variables. The Annals of Applied Statistics 12 2010; **4**(4):2150–2180, DOI:10.1214/10-AOAS355.
25. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 2005; **67**:301–320, DOI:10.1111/j.1467-9868.2005.00503.x.
26. Reulen H. penMSM: Estimating Regularized Multi-state Models Using L1 Penalties. R add-on package published online on the Comprehensive R Archive Network 2015; R package version 0.99.
27. Parmar MS. Chronic renal disease. BMJ 2002; **325**(7355):85–90, DOI:10.1136/bmj.325.7355.85.
28. Allignol A, Beyersmann J, Schumacher M. mvna: An r package for the nelson-aalen estimator in multistate models. R news 2008; **8**(2):48–50.
29. Belitz C, Brezger A, Kneib T, Lang S, Umlauf N. BayesX: Software for Bayesian Inference in Structured Additive Regression Models. Software published online on www.BayesX.org 2012; Version 2.1.
30. Kneib T, Hennerfeind A. Bayesian semi parametric multi-state models. Statistical Modelling 2008; **8**:169–198, DOI:10.1177/1471082X0800800203.
31. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Statistical Science 05 1996; **11**(2):89–121, DOI:10.1214/ss/1038425655.
32. Teixeira L, Cadarso-Suárez C, Rodrigues A, Mendonça D. Assessing the discrimination ability of semiparametric multi-state models in the presence of competing risks. analysis of a peritoneal dialysis program (Unpublished); .
33. Therneau T. survival: A package for survival analysis in s. R add-on package published online on the Comprehensive R Archive Network 2014; R package version 2.37-7.
34. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 2007; **26**(11):2389–2430, DOI:10.1002/sim.2712.
35. Johansen S. An Extension of Cox's Regression Model. International Statistical Review / Revue Internationale de Statistique 1983; **51**(2):165–174, DOI:10.2307/

1402746.

36. Rodríguez-Girondo M, Kneib T, Cadarso-Suárez C, Abu-Assi E. Model building in nonproportional hazard regression. Statistics in Medicine 2013; **32**(30):5301–5314, DOI:10.1002/sim.5961.
37. Carstensen B, Plummer M, Laara E, Hills M. Epi: A Package for Statistical Analysis in Epidemiology. R add-on package published online on the Comprehensive R Archive Network 2014; R package version 1.1.67.
38. Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. Statistics in Medicine 2005; **24**(24):3871–3885, DOI:10.1002/sim.2399.
39. Carstensen B, Center SD. Demography and epidemiology: Practical use of the lexis diagram in the computer age. Annual meeting of Finnish Statistical Society, vol. 23, 2005; 24.

Appendix A. Event History Observations in the Long Format

Throughout this article, the transitions of a multi-state model are denoted by $q = 1, \dots, Q$, and a transition q is composed by an entry state $\text{state}_{\text{entry},q}$ and an exit state $\text{state}_{\text{exit},q}$. Furthermore, entry time $t_{\text{entry},i}$ and exit time $t_{\text{exit},i}$ denote transition (or censoring) times, where the information about non-censoring of the event q is captured by a transition-specific non-censoring indicator $\delta_{q,i}$. In combination, the difference $t_{\text{exit},i} - t_{\text{entry},i}$ between these two time points measures the length of duration of the at-risk spell i. It is important to note that an at-risk spell i here denotes one time interval for one observation unit and each i is represented by a corresponding line in the final dataset, a consequence of using the long format for multi-state model observations as introduced by [34]. The number of competing at-risk processes is visualized by arrows pointing away from one node in the state-chart of a multi-state model, as e.g. in Figures 1 and 2. For the respective entry state $\text{state}_{\text{entry}}$ of a specific observed (or censored) transition, the number of competing at-risk processes defines the number of at-risk spells that result for this (censored) observation.

For example, for a competing risk setting with four competing exit states, one observed event or censoring time leads to four lines in the long format dataset, representing four at-risk spells. For the comprehensive non-censoring indicator δ used in the long data format, only one line out of four has the capability to take the value 1 if it has been actually observed ($\delta = 0$ in case of censoring), the other three out of four lines strictly take value $\delta = 0$. In the exemplary illness-death model with recovery (state-chart in Figure 1) and one time-constant covariate x_p, two exemplary original event history observations with the three states healthy, illness, and death denoted by $\{\text{H}, \text{I}, \text{D}\}$, transitions by $q \in \{\text{HI}, \text{HD}, \text{IH}, \text{ID}\}$, might look like this:

patient id	$\text{state}_{\text{entry}}$	$\text{state}_{\text{exit}}$	t_{entry}	t_{exit}	δ_{HI}	δ_{HD}	δ_{IH}	δ_{ID}	x_p
1	H	D	0	2.28	0	1	0	0	$x_{p,1}$
2	H	I	0	1.5	1	0	0	0	$x_{p,2}$
2	I	H	1.5	7.89	0	0	1	0	$x_{p,2}$
2	H	I	7.89	9.15	1	0	0	0	$x_{p,2}$
2	I	NA	9.15	10	0	0	0	0	$x_{p,2}$

Here, we use transition-specific non-censoring indicators δ_q. The last observation of patient 2 has been right-censored at time 10 and the exit state is therefore not available (NA).

In the long format, these observations lead to the following exemplary data set, where we use a global non-censoring indicator δ and transition-specific covariates $x_{p.\text{HI}}$, $x_{p.\text{HD}}$, $x_{p.\text{IH}}$, and $x_{p.\text{ID}}$:

patient id	$\text{state}_{\text{entry}}$	$\text{state}_{\text{exit}}$	t_{entry}	t_{exit}	δ	$x_{p.\text{HI}}$	$x_{p.\text{HD}}$	$x_{p.\text{IH}}$	$x_{p.\text{ID}}$
1	H	I	0	2.28	0	$x_{p,1}$	0	0	0
1	H	D	0	2.28	1	0	$x_{p,1}$	0	0
2	H	I	0	1.5	1	$x_{p,2}$	0	0	0
2	H	D	0	1.5	0	0	$x_{p,2}$	0	0
2	I	H	1.5	7.89	1	0	0	$x_{p,2}$	0
2	I	D	1.5	7.89	0	0	0	0	$x_{p,2}$
2	H	I	7.89	9.15	1	$x_{p,2}$	0	0	0
2	H	D	7.89	9.15	0	0	$x_{p,2}$	0	0
2	I	H	9.15	10	0	0	0	$x_{p,2}$	0
2	I	D	9.15	10	0	0	0	0	$x_{p,2}$

A data-set in the long-data format will consist of the extended number of n lines, with $n > N$.

The construction of transition-specific covariate vector versions was performed using transition indicators $\psi_{q,i} := \mathrm{I}_{\{\text{state}_{\text{entry},i}=\text{state}_{\text{entry},q}\}}\mathrm{I}_{\{\text{state}_{\text{exit},i}=\text{state}_{\text{exit},q}\}}$. It is now convenient to formulate a general linear predictor

$$\eta_i = (\psi_{1,i}x_{1,i}, \psi_{1,i}x_{2,i}, \dots, \psi_{1,i}x_{P,i}, \psi_{2,i}x_{1,i}, \dots, \psi_{Q,i}x_{P,i},) \cdot \begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \vdots \\ \beta_{P,1} \\ \beta_{1,2} \\ \vdots \\ \beta_{P,Q} \end{pmatrix},$$

in the spirit of general regression models which is inserted in each transition-specific hazard rate formulation in the same way:

$$\lambda_{q,i}\left(t\right) = \lambda_{q,0}\left(t\right)\exp\left(\eta_i\right).$$

Here the index p, $p = 1, \dots, P$, denotes the P covariates used to model the transition-specific factors on the transition-specific baseline hazard rate functions. The allocation of one spell i to its corresponding transition q is described in a broader context in [14] and is a required practice to use the long format. In the following, the vector collecting all transition-specific covariate coefficients will be defined as $\boldsymbol{\beta} := \left(\beta_{1,1}, \dots, \beta_{P,Q}\right)^{\top}$.

Appendix B. Stratified Cox partial likelihood formulation for multi-state models

The Cox log Partial likelihood can be derived as a profile likelihood from the full likelihood (Eq. 1) [35] and is then given in the stratified log Partial likelihood form by

$$\begin{aligned}\mathrm{LogPartialLik}\left(\boldsymbol{\beta}\right) &= \sum_{i=1}^{N}\sum_{q=1}^{Q}\sum_{c=1}^{C_{q,i}(t_{\max,i})} \log\left(\left(\frac{\exp\left(\eta_{q,i}\right)}{\sum\limits_{j=1}^{N}\sum\limits_{c=1}^{C_{q,j}(t_{\max,j})} R_{q,j}\left(t_{q,i,c}\right)\exp\left(\eta_{q,j}\right)}\right)^{\delta_{q,i,c}}\right)\\ &= \sum_{i=1}^{N}\sum_{q=1}^{Q}\sum_{c=1}^{C_{q,i}(t_{\max,i})} \delta_{q,i,c}\left(\eta_{q,i} - \log\left(\sum_{j=1}^{N}\sum_{c=1}^{C_{q,j}(t_{\max,j})} R_{q,j}\left(t_{q,i,c}\right)\exp\left(\eta_{q,j}\right)\right)\right),\end{aligned}$$

where $\eta_{q,i} = \mathbf{x}_i^{\top}\boldsymbol{\beta}_q$. Applying the long format given in Appendix A, we can formulate this in a much more compact way as

$$\mathrm{LogPartialLik}\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n}\delta_i\left(\eta_i - \log\left(\sum_{j\in \mathrm{R}_i}\exp\left(\eta_j\right)\right)\right),$$

where i and j now denote single lines in the long-format data (with n as the number of rows in this data format), which allows us to replace the transition-specific at-risk process $R_{q,j}\left(t_{q,i,c}\right)$ with the risk-set formulation

$$\mathrm{R}_i := \{j : t_{\mathrm{entry},j} < t_{\mathrm{exit},i} \leq t_{\mathrm{exit},j}, \mathrm{state}_{\mathrm{entry},i} = \mathrm{state}_{\mathrm{entry},j}, \mathrm{state}_{\mathrm{exit},i} = \mathrm{state}_{\mathrm{exit},j}\}.$$

This is an at-risk line index notation of the at-risk process described in Section 2.1. The respective information about entry states/times and exit states/times is captured in the long format using vectors $\mathbf{t}_{\mathrm{entry}}, \mathbf{t}_{\mathrm{exit}}, \mathbf{state}_{\mathrm{entry}}, \mathbf{state}_{\mathrm{exit}}$, alike for the event indicator $\boldsymbol{\delta}$ and the matrix of transition-specific covariate information $\mathbf{X}$. The abbreviation $\mathrm{LogPartialLik}\left(\boldsymbol{\beta}\right)$ is used to denote this log partial likelihood in the following.

Appendix C. Piecewise Exponential Model

The piecewise exponential model for single transition survival models with the assumption of piecewise constant baseline hazard rate functions is frequently used in the literature and e.g. described in [36]. The assumption of piecewise constant baseline hazard rate functions requires the definition of an artificial decomposition of the time axis into several sub-intervals as e.g. in Section 2.2. This is called *data augmentation* [36]. There are several possibilities to rely on already available software performing data augmentation, e.g. the function `Lexis` from the R add-on package `Epi` [3, 37]. By this representation of event-times via data augmentation, we are able to use a Poisson maximum likelihood estimation scheme.

For piecewise exponential models, we define a measure $\Delta_{q,i,c}^{(j)}$ that specifies the time-length in the j-th time sub-interval in which individual i was at-risk for the c-th transition of transition q. Here, i denotes one single event-history observation again as in Section 2.1. $\Delta_{q,i,c}^{(j)}$ will be equal to zero in most cases and may take a maximum value of $t^{(j)} - t^{(j-1)}$. Additionally, we define $j_{q,i,c}$ which specifies the sub-interval in which the c-th transition of transition q for individual i occurred. We furthermore include the sub-interval specific constant baseline hazard rate $\lambda_{q,0}^{(j)}$ into the exponential function of the linear predictor by

$$\exp\left(\log\left(\lambda_{q,0}^{(j)}\right) + \mathbf{x}_i^\top \boldsymbol{\beta}\right) =: \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right).$$

Using this, we can now characterize each survival function component by

$$\exp\left(-\sum_{j=1}^{J} \Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right) = \prod_{j=1}^{J} \exp\left(-\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right),$$

and each hazard rate component by

$$\exp\left(\alpha_q^{(j_{q,i,c})} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)^{\delta_{q,i,c}}.$$

The fully composed likelihood is then of the form:

$$\text{Lik} = \prod_{i=1}^{N}\left(\prod_{q=1}^{Q}\left[\prod_{c=1}^{C_{q,i}(T_i)}\left(\exp\left(\alpha_q^{(j_{q,i,c})} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)^{\delta_{q,i,c}} \cdot \prod_{j=1}^{J} \exp\left(-\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right)\right)\right]\right).$$

We now specify sub-interval specific versions $\delta_{q,i,c}^{(j)}$ of $\delta_{q,i,c}$ which take always value zero, despite for the interval $j_{q,i,c}$:

$$\text{Lik} = \prod_{i=1}^{N}\left(\prod_{q=1}^{Q}\left(\prod_{c=1}^{C_{q,i}(T_i)}\left(\prod_{j=1}^{J}\left[\exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)^{\delta_{q,i,c}^{(j)}} \cdot \exp\left(-\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right)\right]\right)\right)\right).$$

Taking the log of this likelihood transforms each of the outer products to a sum with the same indices, and each of the inner factors changes to:

$$\delta_{q,i,c}^{(j)} \cdot \left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right).$$

This yields a likelihood that agrees, except for constants, with the likelihood one would obtain from Poisson distributed observations. For clarification, assume that some δ_i was Poisson distributed with mean $\mu_i = t_i \lambda_i$ and hence results in the following log likelihood contribution (again ignoring additive constants:

$$\begin{aligned} \text{LogLik}_i &= \delta_i \log\left(\mu_i\right) - \mu_i \\ &= \delta_i \log\left(t_i \lambda_i\right) - t_i \lambda_i \\ &= \delta_i \log\left(t_i\right) + \delta_i \log\left(\lambda_i\right) - t_i \lambda_i. \end{aligned}$$

Since $\delta_i \log\left(t_i\right)$ does not depend on any parameter in λ_i, it can be ignored (again a additive constant) from the point of view of estimation:

$$\text{LogLik}_i = \delta_i \log\left(\lambda_i\right) - t_i \lambda_i.$$

We get to a likelihood that is proportional to the likelihood for Poisson distributed response observations $\delta_{q,i,c}^{(j)}$ with mean $\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)$, and so we get to the following log likelihood in the estimation of the piecewise exponential model (n now denotes the number of sub-interval at-risk observations according to the above data augmentation):

$$\text{LogLik}\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \left(-\Delta t_i \cdot \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right) + \delta_i \cdot \log\left(\Delta t_i \cdot \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right)\right),$$

where $\Delta t_i := t_i^{(j)} - t_i^{(j-1)}$ serves as an offset. Conditioning on $\{\mathbf{t}_{\text{entry}}, \mathbf{t}_{\text{exit}}, \mathbf{state}_{\text{entry}}, \mathbf{state}_{\text{exit}}, \boldsymbol{\delta}, \mathbf{X}\}$ is again suppressed for notational simplicity. The linear predictor η_i^{pe} in the piecewise exponential model is composed by:

$$\eta_i^{\text{pe}} = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{q=1}^{Q} \psi_{q,i} \cdot \log\left(f_{t_q}\left(t_{q,i}\right)\right).$$

with $f_{t_q}\left(t_{q,i}\right)$ acting a the transition-specific baseline hazard rate function $\lambda_{q,0}\left(t_i\right) = \exp\left(\alpha_q^{(j)}\right)$, and $\psi_{q,i}$ defined in Appendix A. A specification for modelling continuous covariates that is often used throughout the literature is the class of fractional polynomials [38]. For the estimation of the baseline hazard rate function, we use a specification that was used in the modelling of this type of effect class before [39], i.e.:

$$f_{t_q}\left(t_{q,i}\right) = \sum_{m} t_{q,i}^{m} \beta_{t_q,m},$$

with $m = \left\{\frac{1}{3}, \frac{1}{2}, 0, 1, \frac{3}{2}, 2\right\}$ and $t_{q,i}^{0} := \log\left(t_{q,i}\right)$. This setup is furthermore applicable to any non-linear effect component in the model, as for example the effect of age in the application described in Section 3.

Score vector and Fisher information matrix in the piecewise exponential model For this alternative approach using a piecewise exponential model, the score vector $\mathbf{s}\left(\boldsymbol{\beta}\right) = \dfrac{\partial\,\mathrm{LogLik}\left(\boldsymbol{\beta}\right)}{\partial\,\boldsymbol{\beta}}$ with components $s_p\left(\boldsymbol{\beta}\right) = \dfrac{\partial\,\mathrm{LogLik}\left(\boldsymbol{\beta}\right)}{\partial\,\beta_p}$ is defined by:

$$\mathbf{s}\left(\boldsymbol{\beta}\right) = \mathbf{X}^{\top}\left(\boldsymbol{\delta} - \boldsymbol{\mu}\right),$$

where $\mu := \Delta t_i \cdot \exp\left(\eta_i^{\mathrm{pe}}\right)$. The Fisher information matrix $\mathbf{F}\left(\boldsymbol{\beta}\right) = \dfrac{\partial^2\,\mathrm{LogLik}\left(\boldsymbol{\beta}\right)}{\partial\,\boldsymbol{\beta}\,\partial\,\boldsymbol{\beta}^{\top}}$ with components $F_{p,q}\left(\boldsymbol{\beta}\right) = \dfrac{\partial^2\,\mathrm{LogLik}\left(\boldsymbol{\beta}\right)}{\partial\,\beta_p\,\partial\,\beta_q}$ is defined by

$$\mathbf{F}\left(\boldsymbol{\beta}\right) = \mathbf{X}^{\top}\mathrm{diag}\left(\boldsymbol{\mu}\right)\mathbf{X}.$$

Appendix D. How to set up needed objects in the implemented R package penMSM

A design matrix may be build up in transition-specific blocks of 16 transition-specific covariates, e.g. by the following R [10] command using variables stored in data-set d:

```
X.EP <- as.matrix(d[, c("trans.EP", "bhr.1.EP", "bhr.2.EP", "bhr.3.EP", "bhr.4.EP",
            "bhr.5.EP", "bhr.6.EP", "age.1.EP", "age.2.EP", "age.3.EP", "age.4.EP",
            "age.5.EP", "age.6.EP", "sex.EP", "diab.EP", "prrt.EP")])
X.ED <- as.matrix(d[, c("trans.ED", "bhr.1.ED", ...)])
...
X <- cbind(1, X.EP[, -1], X.ED, X.EH, X.ER, X.PD, X.PH, X.PR)
```

Here, X used for piecewise exponential modelling introduces a constant baseline hazard rate in the first column and defines the transition EP as the reference transition with respect to the constant baseline hazard rate function – which is technically implemented by cbind(1, X.EP[, -1], ...). The resulting design matrix X finally contains 112 columns. Note that the specification of a reference transition, as well as taking into account any baseline hazard rate function columns, is not required for partial likelihood modelling.

The build-up of the penalty structure matrix is conveniently separated into the Lasso part which is caught by matrix PSM1, and the fusion part which is represented by matrix PSM2. The Lasso part penalty structure matrix PSM1 is composed by an identity matrix with the dimension matching the number of columns of X:

```
PSM1 <- diag(ncol(X))
```

Constant baseline hazard rate components equal to 0 seem to be a too restrictive null model for an unbalanced number of transition observations and we therefore leave the constant baseline hazard rates unpenalised:

```
PSM1[1, 1] <- PSM1[17, 17] <- PSM1[33, 33] <- PSM1[49, 49] <-
              PSM1[65, 65] <- PSM1[81, 81] <- PSM1[97, 97] <- 0
```

The penalty structure matrix PSM2 for the fusion part consists of as many columns as the design matrix X and 45 rows, since we want to penalise 15 covariate effects (the constant baseline hazard rates stay again unpenalised) for each of three transition pairs sharing an equal exit state (ED and PD; EH and PH; ER, PR):

```
PSM2 <- matrix(ncol = ncol(X), nrow = 45, 0)
colnames(PSM2) <- colnames(X)
PSM2[ 1, which(colnames(PSM2) %in% c("bhr.1.ED", "bhr.1.PD"))] <- c(-1, 1)
...
PSM2[ 6, which(colnames(PSM2) %in% c("bhr.6.ED", "bhr.6.PD"))] <- c(-1, 1)
PSM2[ 7, which(colnames(PSM2) %in% c("age.1.ED", "age.1.PD"))] <- c(-1, 1)
...
```

```
PSM2[12, which(colnames(PSM2) %in% c("age.6.ED", "age.6.PD"))] <- c(-1, 1)
PSM2[13, which(colnames(PSM2) %in% c("diab.ED", "diab.PD"))] <- c(-1, 1)
PSM2[14, which(colnames(PSM2) %in% c("sex.ED", "sex.PD"))] <- c(-1, 1)
PSM2[15, which(colnames(PSM2) %in% c("prrt.ED", "prrt.PD"))] <- c(-1, 1)
...
PSM2[45, which(colnames(PSM2) %in% c("prrt.ER", "prrt.PR"))] <- c(-1, 1)
```

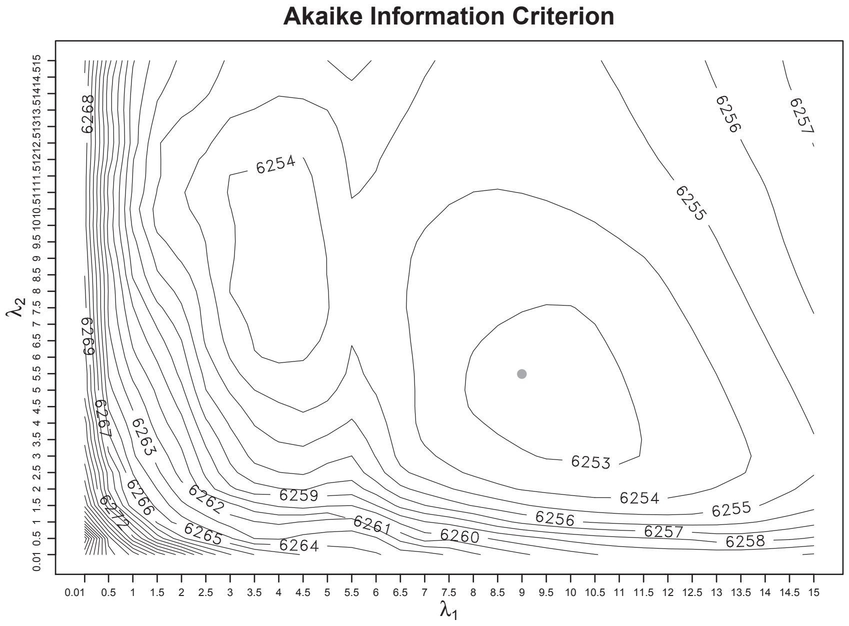

Figure 6: Contour plot of the Akaike Information Criterion values on the (31×31)-grid constructing combinations of the penalty parameters $\lambda_1, \lambda_2 = 0.01, 0.5, 1, 1.5, \ldots, 14.5, 15$. The black lines denote contour lines of the bivariate AIC surface, the grey point indicates the $(\lambda_1, \lambda_2) = (9, 5.5)$ coordinate with the minimal AIC of 6252.54.

Appendix E. Piecewise exponential model results

Using $\Delta_{(j)} = 1$ month is an adequate choice for the length of time sub-intervals since it offers a fine partition of the time axis with observed transition times in the interval $[1, 118]$ on the one hand, and a data-frame with a 44786 lines on the other hand, meaning that the estimation can still be performed in an acceptable time, i.e. a number of 100 iterations of the introduced approach take less than one minute using an ordinary notebook (Intel Core i7 2640M @ 2.80GHz CPU) and the software implementation described in Section 2.6.

We estimate the model on a grid of (λ_1, λ_2) combinations constructed by all pair-wise combinations for $\lambda_1, \lambda_2 = 0.01, 0.5, 1, 1.5, \ldots, 14.5, 15$. The resulting AIC values are illustrated in Figure 6. The minimal AIC value (6252.54) is reached for the combination $(\lambda_1, \lambda_2) = (9, 5.5)$. Any value of (λ_1, λ_2) outside the illustrated region (results not shown) leads to a considerable increase of the AIC, and the minimum at $(9, 5.5)$ therefore stands for the global AIC minimum. The results for the minimum AIC model are illustrated in Figures 8 and 9, and will be described in the following.

Figure 7 shows the paths of the 112 coefficients across the penalty parameter ranges. The left illustration shows the regularisation of the coefficients towards the value 0, the

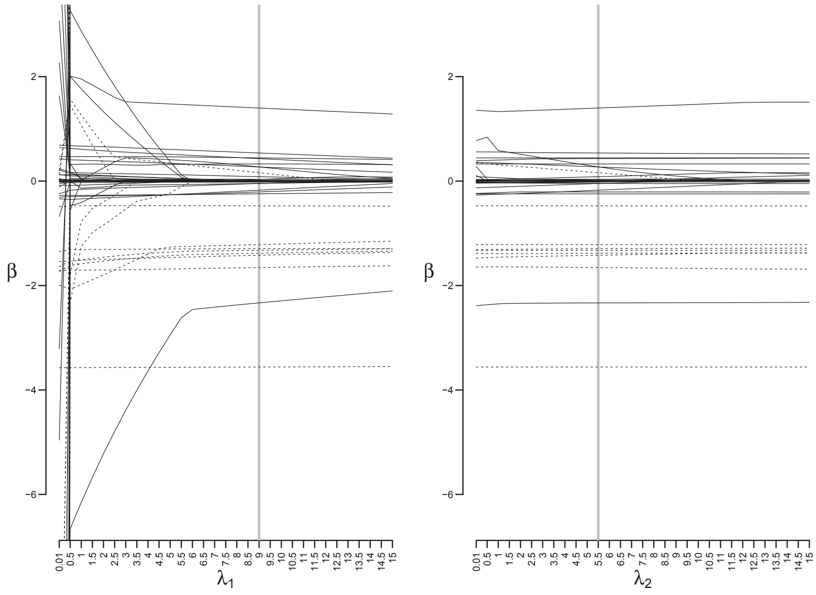

Figure 7: Paths of the regression coefficients in the peritoneal dialysis program data application using piece-wise exponential modelling: for fixed values of the respective other penalty parameter at the value for the minimum AIC model (9 for λ_1 and 5.5 for λ_2, see Figure 6), the penalty parameters λ_1 and λ_2 increase through the values $0.01, 0.5, 1, 1.5, \ldots, 14.5, 15$. Coefficients referring to the log(baseline hazard) are illustrated using dashed lines, coefficients referring to covariate effects are illustrated using solid lines.

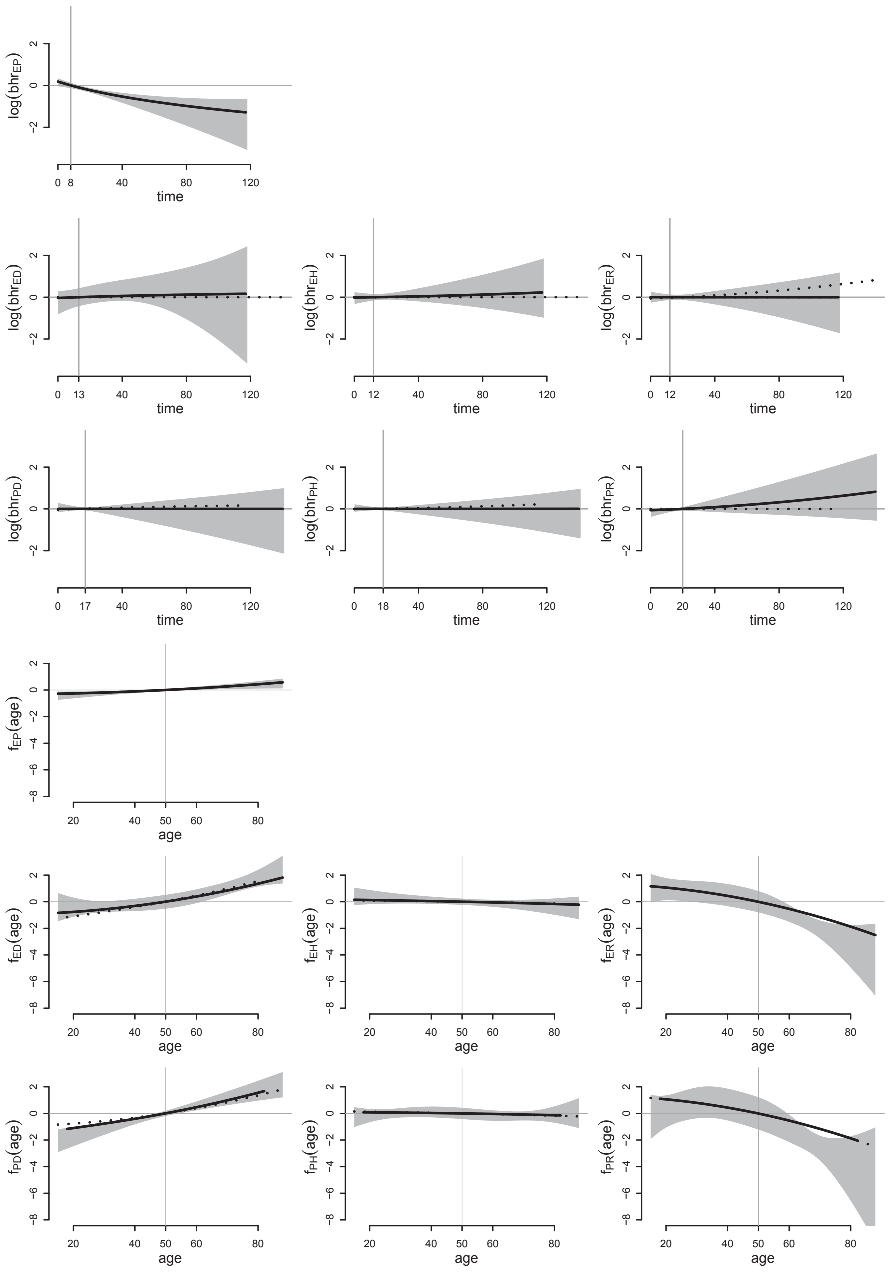

Figure 8: Estimated transition-specific log baseline hazard rate functions (seven plotting windows at the top) and effects of age (seven plotting windows at the bottom) for the piecewise exponential modelling in the peritoneal dialysis program data application with the minimum AIC penalty parameter combination $(\lambda_1, \lambda_2) = (9, 5.5)$: Solid black lines denote the estimated effects by the structured fusion Lasso penalised multi-state model using the fractional-polynomial set-up as described in Appendix E. Dotted lines denote the estimate of the respective fusion penalised effect function (in analogy to the role of the solid and dotted lines in Figure 5). Grey areas illustrate point-wise 95% confidence intervals of the benchmark model using the software BayesX, where smooth effect estimates are obtained using penalised B-splines. The estimated effects and confidence intervals are centred around the median transition-specific transition times for the log baseline hazard rate function estimates and around the values at age 50 for effects of age.

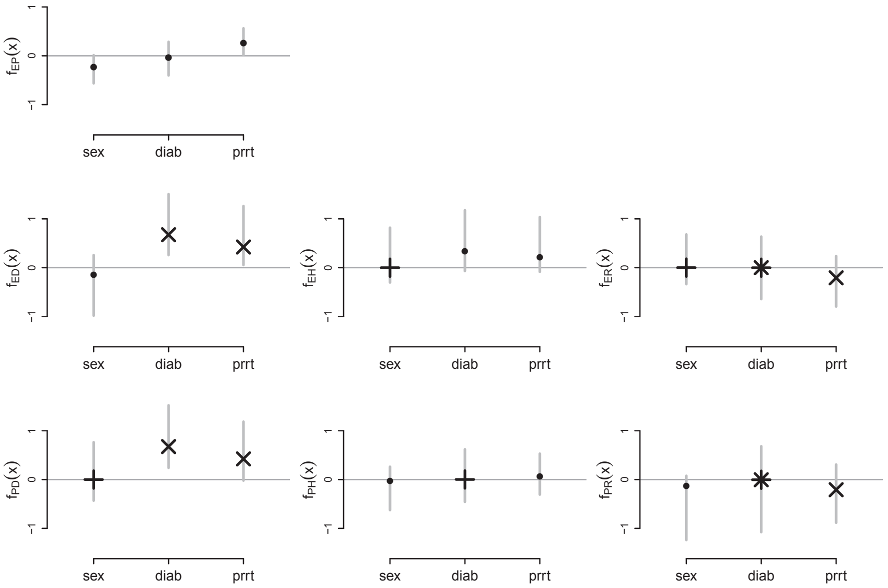

Figure 9: Estimated transition-specific effects of sex, diabetes, and PRRT for the piecewise exponential modelling in the peritoneal dialysis program data application with the minimum AIC penalty parameter combination $(\lambda_1, \lambda_2) = (9, 5.5)$: Black plotting symbols denote the resulting effect by the structured fusion Lasso penalised multi-state model, grey vertical lines illustrate 95% confidence intervals of the benchmark model using the software `BayesX`. Black bullet points ($\bullet$) denote non-fused effects different to 0, plus signs ($+$) denote effects equal to 0, crosses ($\times$) denote fused effects. Combinations of a plus sign and a cross appear as stars ($*$) and signify fused effects equal to 0.

right illustration shows the fusion of coefficients. Both of these regularisation features become stronger with increasing penalty parameters. Note here that, as in any fusion Lasso framework, the influences of the penalty parameters λ_1 and λ_2 on the model coefficients are not independent. For example by changing the Lasso penalty term in most cases different coefficient levels are introduced. Since these changes will be different for different coefficients, the influence of the fusion penalty changes even if the fusion penalty parameter λ_2 is held constant.

As illustrated in Figure 8, we get constant baseline hazard rate function estimates for transitions ER, PD, and PH, but observe no fusion of baseline hazard rate function estimates. However, we receive a cross-transition age effect, i.e. a fusion of transition-specific age effects, for the transition combination ER with PR.

As a result of the structured fusion Lasso penalised estimation we see that the effects for covariate sex are all equal to or smaller than 0. This yields a clearer image of the sex effect in comparison to the un-penalised estimation, where the effect was smaller than 0 for four transitions, and larger than 0 for three other transitions (denoted by the middle of the 95% confidence intervals). To be more precise, all of the effects larger than 0 in the un-penalised estimation are shrunken to 0, whereas all the effects smaller than 0 stay below and different to the value 0. Cross-transition effects are obtained for the transitions to death (for the effect of diabetes and PRRT) and for the transfer to renal transplantation (again for the effects of diabetes and PRRT). Furthermore, the diabetes effect for the transitions ER, PH, and PH are set to 0. Note here that any small coefficient differences preventing setting an effect to 0 or fusing two effects appear also for penalty parameter combinations $(9, 6)$, $(9.5, 5.5)$, and $(9.5, 6)$. Hence they are not attributable to a too coarse grid of penalty parameter combination candidates.