Published in final edited form as:

Biometrics. 2023 September ; 79(3): 1657–1669. doi:10.1111/biom.13761.

Penalized Estimation of Frailty-Based Illness-Death Models for Semi-Competing Risks

Harrison T. Reeder[1,2,*], **Junwei Lu**[3], **Sebastien Haneuse**[3]

[1]Biostatistics, Massachusetts General Hospital, Boston, Massachusetts, U.S.A.

[2]Department of Medicine, Harvard Medical School, Boston, Massachusetts, U.S.A.

[3]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

Summary:

Semi-competing risks refers to the time-to-event analysis setting where the occurrence of a non-terminal event is subject to whether a terminal event has occurred, but not vice versa. Semi-competing risks arise in a broad range of clinical contexts, including studies of preeclampsia, a condition that may arise during pregnancy and for which delivery is a terminal event. Models that acknowledge semi-competing risks enable investigation of relationships between covariates and the joint timing of the outcomes, but methods for model selection and prediction of semi-competing risks in high dimensions are lacking. Moreover, in such settings researchers commonly analyze only a single or composite outcome, losing valuable information and limiting clinical utility—in the obstetric setting, this means ignoring valuable insight into timing of delivery after preeclampsia has onset. To address this gap we propose a novel penalized estimation framework for frailty-based illness-death multi-state modeling of semi-competing risks. Our approach combines non-convex and structured fusion penalization, inducing global sparsity as well as parsimony across submodels. We perform estimation and model selection via a pathwise routine for non-convex optimization, and prove statistical error rate results in this setting. We present a simulation study investigating estimation error and model selection performance, and a comprehensive application of the method to joint risk modeling of preeclampsia and timing of delivery using pregnancy data from an electronic health record.

Keywords

multi-state model; risk prediction; semi-competing risks; structured sparsity; time-to-event analysis; variable selection

[*] hreeder@mgh.harvard.edu .

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 2–7 are available with this paper at the Biometrics website on Wiley Online Library. Software for the proposed methods in R is available at https://github.com/harrisonreeder/SemiCompRisksPen.

1. Introduction

Semi-competing risks refers to the time-to-event analysis setting where a non-terminal event of interest can occur before a terminal event of interest, but not vice versa (Fine et al., 2001). Semi-competing risks are ubiquitous in health research, with example non-terminal events of interest for which death is a semi-competing terminal event including hospital readmission (Lee et al., 2015) and cancer progression (Jazić et al., 2016). An example where death is not the terminal outcome is preeclampsia (PE), a pregnancy-associated hypertensive condition that complicates between 2–8% of all pregnancies and represents a leading cause of maternal and fetal/neonatal mortality and morbidity worldwide (Jeyabalan, 2013). PE can develop during the pregnancy starting at 20 weeks of gestation, but once an individual has given birth they can no longer develop PE, so PE onset and delivery form semi-competing risks. Clinically, the timing of these events is of vital importance. Once PE arises, maternal health risks increase as the pregnancy continues, while giving birth early to alleviate these risks may in turn pose risks to neonatal health and development. Therefore, we are motivated to develop risk models that identify covariates affecting risk and timing of PE while also characterizing the timing of delivery after PE has onset.

Semi-competing risks data represent a unique opportunity to learn about outcomes jointly, by (1) modeling the interplay between the events and baseline covariates, and (2) predicting the covariate-specific risk of experiencing combinations of the outcomes across time. Unfortunately, analysts commonly collapse this joint outcome, considering either the non-terminal or terminal event alone, or a composite endpoint (Jazić et al., 2016). While this enables the use of prediction methods for univariate binary or time-to-event outcomes, modeling risk for one outcome is both a lost opportunity and a severe misalignment with how health-related decisions are actually made; as the PE setting illustrates, clinical care is informed by the joint timing of PE onset and subsequent delivery, not just risk of PE.

Instead, frailty-based illness-death multi-state models (Xu et al., 2010; Lee et al., 2015) characterize the dependency of semi-competing risks and covariates, while also enabling absolute joint risk prediction across time (Putter et al., 2007). These methods comprise three cause-specific hazard submodels for: (i) the non-terminal event; (ii) the terminal event without the non-terminal event; and, (iii) the terminal event after the non-terminal event. Different covariates can affect each hazard differently, and the interplay of these submodels determines the overall covariate-outcome relationship. A person-specific random frailty shared across the submodels captures residual dependence between the two events.

Motivated by application to PE, we consider the task of developing joint risk models for semi-competing risks, specifically in high-dimensional settings such as electronic health records-based studies. Two questions framing model development emerge: (1) which covariates should be included in each submodel, and (2) can information about covariate effects be shared across submodels? To our knowledge only two published papers consider variable selection for these (and related) models, each with important limitations. Sennhenn-Reulen and Kneib (2016) propose ℓ_1-penalized estimation for general multistate models, with parameter-wise penalties inducing sparsity in each submodel and a fused penalty coercing effects for a given covariate to be the same across submodels. This framework, however, does not permit a shared frailty in the model specification, focuses solely on ℓ_1-penalization, and uses a Newton-type algorithm that does not scale to high dimensions. Instead, Chapple et al. (2017) propose a Bayesian spike-and-slab variable selection approach for frailty illness-death models. This framework, however, does not consider linking coefficients across submodels, and is computationally intensive even in low dimensions.

In this paper we propose a novel high-dimensional estimation framework for penalized parametric frailty-based illness-death models. A critical challenge in this setting, however, is that the likelihood-based loss function is non-convex. This renders the development of theoretical results and efficient computational tools particularly difficult. Moreover, to our knowledge no prior literature has examined theoretical properties of penalized frailty-based illness-death models. In relevant work, Loh and Wainwright (2015) prove error bounds for non-convex loss functions with non-convex penalties, but the conditions underlying their result do not directly apply to this setting. Taking into account these various issues, the contributions of this paper are threefold. First, we propose a framework for selecting sparse covariate sets for each submodel via individual non-convex penalties, while inducing parsimony via a fused penalty on effects shared across submodels. Second, we develop a proximal gradient optimization algorithm with a pathwise routine for tuning the model over a grid of regularization parameters. Finally, we prove a high-dimensional statistical error rate for the penalized frailty-based illness-death model estimator. We present a simulation study investigating estimation and model selection properties, and develop a joint risk model for PE and delivery using real pregnancy outcome data from electronic health records.

2. Penalized Illness-Death Model Framework

2.1 Illness-Death Model Specification

Let T_1 and T_2 denote the times to the non-terminal and terminal events, respectively. As outlined in Xu et al. (2010), the illness-death model characterizes the joint distribution of $\mathbf{T} = (T_1, T_2)$ by three hazard functions: a cause-specific hazard for the non-terminal event; a cause-specific hazard for the terminal event in the absence of the non-terminal event; and, a hazard for the terminal event conditional on $T_1 = t_1$. These three hazards can be structured as a function of covariates, denoted $\mathbf{X}$, and an individual-specific random frailty, denoted γ, to add flexibility in the dependence structure between T_1 and T_2, as follows:

$$h_1^c(t_1 \mid \mathbf{X}_1, \gamma) = \lim_{\Delta \downarrow 0} \Delta^{-1} \Pr(T_1 \in [t_1, t_1 + \Delta) \mid T_1 \geqslant t_1, T_2 \geqslant t_1, \mathbf{X}_1, \gamma), \quad t_1 > 0,$$

$$h_2^c(t_2 \mid \mathbf{X}_2, \gamma) = \lim_{\Delta \downarrow 0} \Delta^{-1} \Pr(T_2 \in [t_2, t_2 + \Delta) \mid T_1 \geqslant t_2, T_2 \geqslant t_2, \mathbf{X}_2, \gamma), \quad t_2 > 0,$$

$$h_3^c(t_2 \mid t_1, \mathbf{X}_3, \gamma) = \lim_{\Delta \downarrow 0} \Delta^{-1} \Pr(T_2 \in [t_2, t_2 + \Delta) \mid T_1 = t_1, T_2 \geqslant t_2, \mathbf{X}_3, \gamma), \quad t_2 > t_1 > 0,$$

where $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ are each subsets of $\mathbf{X}$. Practically, in order to make progress, one must specify some form of structure regarding the dependence of each hazard on $\mathbf{X}$ and γ. In this paper we focus on the class of multiplicative-hazard regression models, of the form:

$$h_1^c(t_1 \mid \mathbf{X}_1, \gamma) = \gamma h_1(t_1 \mid \mathbf{X}_1) = \gamma h_{01}(t_1)\exp\left(\mathbf{X}_1^\top \boldsymbol{\beta}_1\right), \quad t_1 > 0, \tag{1}$$

$$h_2^c(t_2 \mid \mathbf{X}_2, \gamma) = \gamma h_2(t_2 \mid \mathbf{X}_2) = \gamma h_{02}(t_2)\exp\left(\mathbf{X}_2^\top \boldsymbol{\beta}_2\right), \quad t_2 > 0, \tag{2}$$

$$h_3^c(t_2 \mid t_1, \mathbf{X}_3, \gamma) = \gamma h_3(t_2 \mid t_1, \mathbf{X}_3) = \gamma h_{03}(t_2 \mid t_1)\exp\left(\mathbf{X}_3^\top \boldsymbol{\beta}_3\right), \quad t_2 > t_1 > 0, \tag{3}$$

where h_{0g} is a transition-specific baseline hazard function, $g = 1, 2, 3$, and $\boldsymbol{\beta}_g \in \mathbb{R}^{d_g}$ is a d_g-vector of transition-specific log-hazard ratio regression coefficients. For ease of notation of the hazard functions, we subsequently suppress conditionality on γ and $\mathbf{X}_g$.

Within this class of models, analysts must make several choices about the structure of the specific model to be adopted. First, it must be decided how exactly the non-terminal event time T_1 affects $h_{03}(t_2 \mid t_1)$ in submodel (3). The so-called Markov structure sets $h_{03}(t_2 \mid t_1) = h_{03}(t_2)$, meaning the baseline hazard is independent of t_1. Alternatively, the semi-Markov structure sets $h_{03}(t_2 \mid t_1) = h_{03}(t_2 - t_1)$, so the time scale for h_3 becomes the time from the non-terminal event to the terminal event (sometimes called the sojourn time) (Putter et al., 2007; Xu et al., 2010). Semi-Markov specification also allows functions of t_1 as covariates in the h_3 submodel, to further capture dependence between T_1 and T_2. While the choice of structure changes the interpretation of h_3 and $\boldsymbol{\beta}_3$, the proposed penalization framework and theory allow either specification. Rather than focus on one or the other, we use semi-Markov specification when writing equations to simplify notation, as well as in the simulation study. However, in our application to pregnancy data we use Markov models to facilitate interpretation of the resulting estimates on the scale of gestational age.

A second important choice concerns the form of the three baseline hazard functions. Given the overarching goals of this paper, we focus on parametric specifications with a fixed-dimensional k_g-vector of unknown parameters, $\boldsymbol{\phi}_g = \left(\phi_{g1}, \ldots, \phi_{gk_g}\right)^\top$, for the gth baseline hazard. For example, one could adopt a form arising from some specific distribution such as the Weibull distribution: $h_{0g}(t) = \exp(\phi_{g1} + \phi_{g2}) \cdot t^{\exp(\phi_{g1}) - 1}$. More flexible options include the piecewise constant baseline hazard defined as $h_{0g}(t) = \sum_{j=1}^{k_g} \exp(\phi_{gj}) \mathbb{1}\left(t^{(j)} \leqslant t < t^{(j+1)}\right)$, with a user-defined set of breakpoints $0 = t^{(1)} < \cdots < t^{(k_g)} < t^{(k_g + 1)} = \infty$. Web Appendix H describes other possible flexible spline-based baseline hazard specifications.

Finally, one must choose a distribution for γ, the individual-specific frailties. These terms serve to capture additional within-subject correlation between T_1 and T_2 beyond covariate effects, and increase flexibility beyond the assumed baseline hazard specification and Markov or semi-Markov model structure (Xu et al., 2010). In this, the frailties play a role that is analogous to that of random effects in generalized linear mixed models. As discussed in Web Appendix B, inclusion of frailties also helps to characterize variability in individualized risk predictions. While in principle one could adopt any distribution for γ, we focus on the common choice of $\gamma \sim \text{Gamma}(e^{-\sigma}, e^{-\sigma})$. This distribution has mean 1 and variance e^{σ}, and uniquely yields a closed form marginal likelihood, as shown in (4). This log-variance parameter σ can be interpreted as characterizing residual variability of the outcomes beyond the specified baseline hazard and transition model structure.

2.2 The Observed Data Likelihood

Now, let C denote the right-censoring time. The observable outcome data for the ith subject is then $\mathscr{D}_i = \{Y_{1i}, \Delta_{1i}, Y_{2i}, \Delta_{2i}, \mathbf{X}_i\}$ where $Y_{1i} = \min(C_i, T_{1i}, T_{2i})$, $Y_{2i} = \min(C_i, T_{2i})$, $\Delta_{1i} = \mathbb{I}(Y_{1i} = T_{1i})$, and $\Delta_{2i} = \mathbb{I}(Y_{2i} = T_{2i})$. We denote the corresponding observed outcome values as y_{1i}, y_{2i}, δ_{1i}, and δ_{2i}. Given specification of models (1), (2) and (3), let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^{\top}, \boldsymbol{\phi}_2^{\top}, \boldsymbol{\phi}_3^{\top})^{\top}$ denote the k×1 vector of baseline hazard components, with $k = k_1 + k_2 + k_3$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\top}, \boldsymbol{\beta}_2^{\top}, \boldsymbol{\beta}_3^{\top})^{\top}$ the $d \times 1$ vector of log-hazard ratios, with $d = d_1 + d_2 + d_3$. Finally, let $\boldsymbol{\psi} = (\boldsymbol{\beta}^{\top}, \boldsymbol{\phi}^{\top}, \sigma)^{\top}$ denote the full set of $d + k + 1$ unknown parameters.

To develop the observed data likelihood, we assume independencies between: frailty and covariates, $\gamma \perp\!\!\!\perp \mathbf{X}$; frailty and censoring time given covariates, $\gamma \perp\!\!\!\perp C \mid \mathbf{X}$; and, censoring time and event times, given covariates and frailty, $C \perp\!\!\!\perp \mathbf{T} \mid (\gamma, \mathbf{X})$. Illustrating under semi-Markov specification, the ith likelihood contribution given γ_i is $\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\phi} \mid \gamma_i, \mathscr{D}_i) = h_1^c(y_{1i})^{\delta_{1i}} h_2^c(y_{1i})^{(1-\delta_{1i})\delta_{2i}} h_3^c(y_{2i} - y_{1i})^{\delta_{1i}\delta_{2i}} \exp\{-H_1^c(y_{1i}) - H_2^c(y_{1i}) - \delta_{1i} H_3^c(y_{2i} - y_{1i})\}$, where $H_g^c(t) = \int_0^t h_g^c(s)ds$. Finally, integrating out the gamma-distributed frailty, the ith marginal likelihood contribution takes the closed form

$$\begin{aligned} \mathscr{L}(\boldsymbol{\psi} \mid \mathscr{D}_i) &= h_1(y_{1i})^{\delta_{1i}} h_2(y_{1i})^{(1-\delta_{1i})\delta_{2i}} h_3(y_{2i} - y_{1i})^{\delta_{1i}\delta_{2i}} (1 + e^{\sigma})^{\delta_{1i}\delta_{2i}} \\ &\times (1 + e^{\sigma}\{H_1(y_{1i}) + H_2(y_{1i}) + H_3(y_{2i} - y_{1i})\})^{-\exp(-\sigma) - \delta_{1i} - \delta_{2i}}. \end{aligned} \tag{4}$$

2.3 Penalization for Sparsity and Model Parsimony

Given an i.i.d sample of size n, let $\ell(\boldsymbol{\psi}) = -n^{-1}\sum_{i=1}^{n} \log \mathscr{L}(\boldsymbol{\psi} \mid \mathscr{D}_i)$ denote the negative log-likelihood. Penalized likelihood estimation follows via the introduction of a penalty function $P_{\lambda}(\boldsymbol{\psi})$, yielding a new objective function of the form $Q_{\lambda}(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + P_{\lambda}(\boldsymbol{\psi})$.

Letting $\mathbf{X} = \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X}_3$, the illness-death model's hazards allow each X_j to potentially have three different coefficients: a cause-specific log-hazard ratio for the non-terminal event ($\boldsymbol{\beta}_{1j}$), a cause-specific log-hazard ratio for the terminal event ($\boldsymbol{\beta}_{2j}$), and a log-hazard ratio for the terminal event given the non-terminal event has occurred ($\boldsymbol{\beta}_{3j}$). We propose a structured $P_{\lambda}(\boldsymbol{\psi})$ simultaneously targeting two properties: (i) sparsity, by identifying important non-zero covariate effects, and (ii) parsimony, by identifying relationships between the effects of each covariate across the three submodels. We propose the general form

$$P_{\lambda}(\boldsymbol{\psi}) = \sum_{g=1}^{3} \sum_{j=1}^{d_g} p_{\lambda_1}(|\beta_{gj}|) + \sum_{g \neq g'} \sum_{j=1}^{d_g} p_{\lambda_2}(|\beta_{gj} - \beta_{g'j}|). \tag{5}$$

The first component, regulated by λ_1, induces sparsity by setting unimportant covariate effects to zero. The second component, regulated by λ_2, induces parsimony by regularizing the cofficients of each covariate X_j towards being similar or shared across submodels.

For each component, one could, in principle, consider any of a wide array of well-known penalties, such as the Lasso ℓ_1 penalty (Tibshirani, 1996) or non-convex penalties like smoothly-clipped absolute deviation (SCAD) (Fan and Li, 2001):

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & |\beta| \leqslant \lambda, \\ -\left\{\beta^2 - 2\xi\lambda\left|\beta\right| + \lambda^2\right\}/\left\{2(\xi - 1)\right\}, & \lambda < |\beta| \leqslant \xi\lambda, \\ (\xi + 1)\lambda^2/2, & |\beta| > \xi\lambda, \end{cases} \tag{6}$$

with $\xi > 2$ controlling the level of non-convexity. This penalty behaves like the Lasso near zero, but flattens out for larger values, reducing bias on truly non-zero estimates.

The motivation for this form of penalty is both clinical and statistical, and we emphasize that depending on the application, the fusion penalties between $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and/or $\boldsymbol{\beta}_3$ can be included or omitted from $P_\lambda(\boldsymbol{\psi})$. Clinically, fusion penalties are valuable when there is subject matter knowledge indicating that covariates likely have similar effects in two or more submodels. For example, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ both represent log-hazard ratios for the terminal event, with $\boldsymbol{\beta}_2$ representing cause-specific effects in the absence of the non-terminal event, and $\boldsymbol{\beta}_3$ representing effects conditional on the occurrence of the non-terminal event. Therefore, fusing $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ imposes a structure where covariate effects on the terminal event are similar whether or not the non-terminal event has occurred. Relatedly, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ both represent cause-specific log-hazard ratios, for the non-terminal and terminal event respectively. Therefore, in settings where both non-terminal and terminal events represent negative health outcomes, like cancer progression and death, fusing these components induces each covariate to have similar or shared cause-specific hazard ratio estimates for the two events. In any case, well-chosen structured fusion penalties can be used to encode clinically meaningful subject matter knowledge into the estimation framework.

Statistically, fusion penalties may also be valuable when there is relatively little information on one of these submodels, and the goal is to impose structure and stabilize estimation. For example, in settings where the non-terminal event is rare relative to the terminal event, there will be more information available for estimating $\boldsymbol{\beta}_2$ relative to $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_3$. Therefore, adding a fusion penalty between $\boldsymbol{\beta}_2$ and either $\boldsymbol{\beta}_1$ and/or $\boldsymbol{\beta}_3$ regularizes the more variable estimates of $\boldsymbol{\beta}_1$ and/or $\boldsymbol{\beta}_3$ towards the more precise estimates of $\boldsymbol{\beta}_2$, effectively borrowing information across submodels. As with all regularized estimation, this directly reflects a bias-variance trade off: imposing structure on covariate effects across hazards to reduce variance, or leaving effects unstructured across hazards to reduce bias.

3. Optimization

Practically, minimizing the objective function $Q_\lambda(\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ poses several interconnected challenges: the loss function ℓ is non-convex due to the marginalized random frailty; the penalty functions p_λ may also be non-convex; and, the fusion penalty component does not admit standard algorithms for general fused Lasso tailored to linear regression (Tibshirani and Taylor, 2011). Finally, the combination of penalties requires tuning a two-dimensional regularization parameter $\boldsymbol{\lambda}$. In this section, we propose a comprehensive optimization routine to simultaneously and efficiently handle these challenges.

3.1 Proximal Gradient Descent with a Smoothed Fusion Penalty

Proximal gradient descent iteratively minimizes objective functions like Q_λ defined as the sum of a differentiable loss function and a non-differentiable penalty. When $P_\lambda(\boldsymbol{\psi})$ is the standard Lasso ℓ_1-penalty $\lambda\|\boldsymbol{\beta}\|_1$, the algorithm reduces to standard gradient descent with an added soft-thresholding operation. To leverage this property, we combine two techniques to recast Q_λ from a loss with a complex penalty into a loss with a simple Lasso penalty.

First we decompose each p_λ in (5) into the sum of a smooth concave term and a simple ℓ_1 penalty term, of the form $p_\lambda(|x|) = \tilde{p}_\lambda(|\beta|) + \lambda|\beta|$, where $\tilde{p}_\lambda$ is a Lipschitz-smooth concave function (Zhao et al., 2018). The goal is to treat the smooth component as part of the likelihood, leaving only a simpler ℓ_1 penalty. This decomposition can be done to both the parameterwise penalties p_{λ_1} and the fusion penalties p_{λ_2}, rewriting (5) as

$$P_\lambda(\boldsymbol{\psi}) = \sum_{g=1}^{3}\sum_{j=1}^{d_g} \tilde{p}_{\lambda_1}(|\beta_{gj}|) + \sum_{g \neq g'}\sum_{j=1}^{d_g} \tilde{p}_{\lambda_2}(|\beta_{gj} - \beta_{g'j}|) + \lambda_1\|\boldsymbol{\beta}\|_1 + \Omega_{\lambda_2}(\boldsymbol{\psi}), \tag{7}$$

where $\Omega_{\lambda_2}(\boldsymbol{\psi}) = \sum_{g \neq g'} \sum_{j=1}^{d_g} \lambda_2|\beta_{gj} - \beta_{g'j}|$ denotes the fusion ℓ_1 penalty.

However, this fusion penalty $\Omega_{\lambda_2}(\boldsymbol{\psi})$ still complicates optimization, so next we use Nesterov smoothing to substitute it with a smoothed, differentiable surrogate (Chen et al., 2012). Defining $\mathbf{D}_{\lambda_2}$ as a contrast matrix such that $\Omega_{\lambda_2}(\boldsymbol{\psi}) = \|\mathbf{D}_{\lambda_2}\boldsymbol{\psi}\|_1$, the surrogate is

$$\widetilde{\Omega}_{\lambda_2,\mu}(\boldsymbol{\psi}) = \max_{\|\mathbf{z}\|_\infty \leqslant 1}\left(\mathbf{z}^\top \mathbf{D}_{\lambda_2}\boldsymbol{\psi} - \mu\|\mathbf{z}\|_2^2/2\right) = (\mathbf{z}^*)^\top \mathbf{D}_{\lambda_2}\boldsymbol{\psi} - \mu\|\mathbf{z}^*\|_2^2/2, \tag{8}$$

where $\mathbf{z}^* = \mathbf{S}(\mathbf{D}_{\lambda_2}\boldsymbol{\psi}/\mu)$ and $\mathbf{S}(\mathbf{x})$ is the vector-valued projection operation onto the unit box, defined at the jth element by $[\mathbf{S}(\mathbf{x})]_j = \text{sign}(x_j)\max(1, |x_j|)$ and $\mu > 0$ is a user-chosen smoothness parameter. Smaller μ yields a tighter approximation, with the gap between penalty and surrogate bounded by $\Omega_{\lambda_2}(\boldsymbol{\psi}) - \mu J/2 \leqslant \widetilde{\Omega}_{\lambda_2,\mu}(\boldsymbol{\psi}) \leqslant \Omega_{\lambda_2}(\boldsymbol{\psi})$, where J is the number of pairwise fusion terms. Web Appendix G details tuning methods for μ.

Together, (7) and (8) recast the objective function $Q_\lambda(\boldsymbol{\psi})$ as an ℓ_1-penalized objective:

$$Q_{\lambda,\mu}(\boldsymbol{\psi}) = \widetilde{\ell}_{\lambda,\mu}(\boldsymbol{\psi}) + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{9}$$

where $\widetilde{\ell}_{\lambda,\mu}(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + \sum_{g=1}^{3}\sum_{j=1}^{d_g}\tilde{p}_{\lambda_1}(|\beta_{gj}|) + \sum_{g \neq g'}\sum_{j=1}^{d_g}\tilde{p}_{\lambda_2}(|\beta_{gj} - \beta_{g'j}|) + \widetilde{\Omega}_{\lambda_2,\mu}(\boldsymbol{\psi})$. Towards optimizing (9), define the vector-valued soft thresholding operation $\mathbf{S}_\lambda(\mathbf{x})$ at the jth element by $[\mathbf{S}_\lambda(\mathbf{x})]_j = \text{sign}(x_j)\max(0, |x_j| - \lambda)$. Then the mth step of the iterative proximal gradient algorithm is $\boldsymbol{\psi}^{(m)} \leftarrow \mathbf{S}_{\lambda_1}\left\{\boldsymbol{\psi}^{(m-1)} - r^{(m)} \cdot \nabla\widetilde{\ell}_{\lambda,\mu}\left(\boldsymbol{\psi}^{(m-1)}\right)\right\}$, where $r^{(m)}$ is an adaptive step size determined by backtracking line search (see, e.g., ‘Algorithm 3’ of Wang et al., 2014).

Iterations continue until change in objective function $|Q_{\lambda,\mu}(\boldsymbol{\psi}^{(m)})-Q_{\lambda,\mu}(\boldsymbol{\psi}^{(m-1)})|$ falls below a given threshold (e.g., 10^{-6} in this paper).

3.2 Tuning Regularization Parameters via Pathwise Grid Search

For non-convex penalized problems with a single regularization parameter λ, recent path-following routines apply proximal gradient descent or coordinate descent over a decreasing sequence of regularization parameters (Wang et al., 2014; Zhao et al., 2018). At each new λ, these routines initialize at the solution of the prior λ. The result is a sequence of estimates across a range of penalization levels, also called a regularization path. Under certain conditions these ‘approximate path-following’ approaches yield high-quality local solutions with attractive theoretical properties, even when the loss and/or regularizer are non-convex. Heuristically, many non-convex objective functions are locally convex in the neighborhood of well-behaved optima, and so incrementally optimizing over a sequence of small changes to λ ensures that each local solution remains in the convex neighborhood of the solution under the previous λ.

Therefore, we develop a pathwise approach to the penalized illness-death model (9) with a search routine over a two-dimensional grid of the sparsity parameter λ_1 and fusion parameter λ_2 (Figure 1). This routine consists of an outer loop decrementing λ_1 as in standard pathwise algorithms, and an inner loop comprising a branching pathwise search over increasing λ_2 values. The resulting grid search of the model space slowly grows the number of non-zero coefficient estimates as λ_1 decreases in the outer loop, and then explores how the resulting non-zero coefficients fuse as λ_2 increases in the inner loop. Assuming sparsity of the regression coefficients, a straightforward choice to begin the pathwise regularization grid search routine is to set $\boldsymbol{\beta} = \mathbf{0}$, and set the remaining parameters to the unadjusted MLE estimates fit without covariates. Optimization at each point initializes from the solution at the prior relevant step.

The grid should be as fine as computational costs allow; Wang et al. (2014) recommend that successive values of λ_1 differ by no more than a factor of 0.9, and for λ_2 we chose four grid points in simulations and seven for the data application. Final choice of (λ_1, λ_2) follows by minimizing a performance metric computed at each grid point, depicted by shading at each point in Figure 1. Metrics such as Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) may be computed using model degrees of freedom estimated by the number of unique covariate estimates (Sennhenn-Reulen and Kneib, 2016).

4. Theoretical Results

In this section, we derive the statistical error rate for estimation of the true parameter vector $\boldsymbol{\psi}^*$ in a gamma frailty illness-death model with non-convex penalty, encompassing high-dimensional settings where $d > n$ with sparsity level denoted $\|\boldsymbol{\psi}^*\|_0 = s$. This work builds on the framework of Loh and Wainwright (2015), extended to the additional complexities of parametric gamma-frailty illness-death models. We develop a set of sufficient conditions for this setting under which we prove the statistical rate, and verify that such conditions hold with high probability under several common model specifications.

This statistical investigation focuses on the estimator

$$\widehat{\boldsymbol{\psi}} = \underset{\|\boldsymbol{\psi}\|_1 \leqslant R_1}{\arg\min}\, Q_\lambda(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + \sum_{j=1}^{d_1} p_\lambda(|\beta_{1j}|) + \sum_{g=2}^{3} \sum_{j=1}^{d_g} p_\lambda(|\beta_{gj} - \beta_{1j}|)\,. \tag{10}$$

We note that while the general framework (5) allows a penalty on every element and pairwise difference of the regression parameter vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$, the estimator based on (10) specifically penalizes $\boldsymbol{\beta}_1$ and its pairwise differences with $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$. This facilitates the theoretical analysis while retaining the property that the elementwise differences between $\boldsymbol{\beta}_g$'s are sparse. Moreover, to accommodate the role of non-convexity the constraint $\|\boldsymbol{\psi}\|_1 \leqslant R_1$ is imposed on the parameter space over which solutions are sought (Loh and Wainwright, 2015). See Assumption 2 below and its remark for detailed discussion.

We start by listing assumptions used to derive the statistical rate of the estimator of $\boldsymbol{\psi}^*$ based on (10). To unify outcome notation across all three submodels $g = 1, 2, 3$, let

$$\widetilde{Y}_{gi} = \begin{cases} Y_{1i}, & g = 1, 2, \\ Y_{2i} - Y_{1i}, & g = 3, \end{cases} \quad \text{and} \quad \widetilde{\Delta}_{gi} = \begin{cases} \Delta_{1i}, & g = 1, \\ (1 - \Delta_{1i})\Delta_{2i}, & g = 2, \\ \Delta_{1i}\Delta_{2i}, & g = 3, \end{cases}$$

where Y_{gi} and Δ_{gi} are defined as in Section 2.2.

Assumption 1 (Bounded Data): There exists some administrative maximum time τ_Y such that $0 < Y_1 \leqslant Y_2 \leqslant \tau_Y < \infty$. Additionally, there exists some positive covariate bound τ_X such that $\tau_X \geqslant \|\mathbf{X}\|_\infty$, where $\|\mathbf{X}\|_\infty = \max_{j=1,\ldots,d}|X_j|$.

Assumption 1 ensures boundedness of the observed data. This assumption will invariably be satisfied in real world data applications, especially in time-to-event studies where person-time is censored and there are practical limits on covariate values.

Assumption 2 (Bounded True Parameter): There exists a $R_2 > 0$ such that $\|\boldsymbol{\psi}^*\|_1 \leqslant R_2$.

Assumption 2 characterizes the overall length of the true parameter vector in terms of ℓ_1-norm. Combined with the side constraint introduced in (10) and setting $R = R_1 + R_2$, this ensures by the triangle inequality that there is an overall bound $\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|_1 \leqslant R$ for each iterate and all stationary points of the optimization routine.

Assumption 3 (Bounded Minimum Population Hessian Eigenvalue): There exists a $\rho > 0$ such that $\min_{\boldsymbol{\psi}: \|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|_2 \leqslant R} \lambda_{\min}\{\sum(\boldsymbol{\psi})\} \geqslant \rho$, where $\lambda_{\min}\{\Sigma(\boldsymbol{\psi})\}$ is the minimum eigenvalue of $\Sigma(\boldsymbol{\psi}) = \mathbb{E}\left\{\nabla^2 \ell(\boldsymbol{\psi})\right\}$, the population Hessian matrix at $\boldsymbol{\psi}$.

Assumption 3 characterizes the positive-definiteness of the expected Hessian matrix of the loss function as a function of $\boldsymbol{\psi}$, and guarantees curvature of the population loss function in a neighborhood around the truth.

Assumption 4 (Baseline Hazard Function Sufficient Conditions): For g, $r = 1, 2, 3$, $j = 1, \ldots, k_g$, and $l = 1, \ldots, k_r$, and for all $\{\boldsymbol{\psi} : \|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|_2 \leqslant R\}$,

a. $H_{0g}(t)$, $\partial H_{0g}(t)/\partial \phi_{gj}$, and $\partial^2 H_{0g}(t)/(\partial \phi_{gj}\partial \phi_{rl})$ are bounded functions on $0 \leqslant t \leqslant \tau$ for any $\tau > 0$.

b. $\mathrm{Var}\left\{\tilde{\Delta}_{gi}\partial \log h_{0g}\left(\tilde{Y}_{gi}\right)/\partial \phi_{gj}\right\}$ is finite.

c. Each log-hazard second derivative factorizes into the form $\partial^2 \log h_{0g}(t)/(\partial \phi_{gj}\partial \phi_{rl}) = w_{jl}^{gr}(\boldsymbol{\psi})z_{jl}^{gr}(t)$, where $w_{jl}^{gr}(\boldsymbol{\psi})$ is only a function of $\boldsymbol{\psi}$ and $z_{jl}^{gr}(t)$ is only a function of t. In addition, every $\mathrm{Var}\left\{\tilde{\Delta}_{gi}z_{jl}^{gr}\left(\tilde{Y}_{gi}\right)\right\}$ is finite.

Assumption 4 outlines conditions regarding the baseline hazard functions in the illness-death model specification. Collectively, these conditions are imposed to control the maximum deviations of the gradient $\|\nabla \ell(\boldsymbol{\psi}^*)\|_\infty$, and Hessian $\|\nabla^2 \ell(\boldsymbol{\psi}) - \Sigma(\boldsymbol{\psi})\|_{\max}$ for all $\boldsymbol{\psi}$ over the ℓ_2-ball $\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|_2 \leqslant R$, where $\|\cdot\|_{\max}$ is the matrix elementwise absolute maximum. Specifically, the gradient and Hessian of the empirical loss function ℓ may be unbounded, which complicates our analysis; under a Weibull specification, for example, several elements of the gradient $\nabla \ell(\boldsymbol{\psi})$ involve the term $\log \tilde{y}_{gi}$, which diverges approaching 0. As such, Assumptions 4b and 4c are used to control the unbounded quantities, while Assumption 4a bounds remaining terms. Note, these conditions are satisfied by commonly-used baseline hazard choices; Web Appendix E contains proofs for piecewise constant and Weibull specifications. Lastly, while the conditions in Assumption 4 are presented under a semi-Markov model, analogous conditions can be expressed for a Markov model.

We now present the main theorem on the statistical rate of the estimator in (10). We take p_λ to be the SCAD penalty defined in (6) to streamline the statement in terms of SCAD's non-convexity parameter ξ, though the result holds for other penalty functions including the Lasso and minimax concave penalty (MCP) (Zhang, 2010), as described in Web Appendix C.

THEOREM 1: *Under* Assumptions 1, 2, *and* 3 *and sparsity level* $s = \|\boldsymbol{\psi}^*\|_0$, *consider a gamma frailty illness-death model satisfying* Assumption 4 *with SCAD penalization as in* (10). *Suppose the SCAD non-convexity parameter* ξ *satisfies* $3/\{4(\xi - 1)\} < \rho$, *where* ρ *is the population Hessian eigenvalue bound defined in* Assumption 3. *Then choosing* $\lambda = c\sqrt{\log(dn)/n}$ *for sample size* n, *parameter dimensionality* d, *and sufficiently large constant* c, *any stationary point* $\hat{\boldsymbol{\psi}}$ *of* (10) *will have a statistical rate that varies with* s, n, *and* d *as*

$$\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*\|_2 = O_P(\sqrt{s\log(dn)/n}).$$

The proof of this theorem and detailed discussion are left to Web Appendix C. In particular, due to the complexities outlined in the discussion of Assumption 4, the proof relies on a weaker version of the so-called Restricted Strong Convexity condition than that of Loh and Wainwright (2015). Lastly, we note that by this result, consistency of the estimator $\hat{\boldsymbol{\psi}}$ follows in the high-dimensional regime under scaling condition $s\log(dn)/n \to 0$.

5. Simulation Studies

In this section, we present a series of simulation studies to investigate the performance of the proposed methods in terms of estimation error and selection of the covariate effects $\boldsymbol{\beta}$, comparing various penalty specifications with ad hoc methods like forward selection.

5.1 Set-up and Data Generation

We consider eight simulation scenarios, each based on a true semi-Markov illness death model with gamma frailty variance $e^{\sigma} = 0.5$. The eight scenarios arise as all combinations of two specifications for each of the baseline hazard functions, the overall covariate dimensionality and the values of the true regression parameters. The specifications under consideration are detailed in Web Table F.1, and summarized below. We repeated simulations under the given settings for three sample sizes: $n = 300, 500, 1000$.

The true baseline hazard specifications were piecewise constant with breakpoints at 5, 15, and 20, specified to yield particular marginal event rates for the non-terminal event. Under the ‘Low Non-Terminal Event Rate’ setting approximately 17% of subjects are observed to experience the non-terminal event, while under the ‘Moderate Non-Terminal Event Rate’ setting this number was 30%. Both specifications represent complex non-monotonic hazards not be well-approximated by Weibull parameterization, to examine the impact of such misspecification on regression parameter selection and estimation error.

We considered both low- and high-dimensional regimes under sparsity, with 25 and 350 covariates respectively, having 10 true non-zero coefficients in each submodel ranging in magnitude from 0.2 to 1. Crucially, the high-dimensional setting always has more regression parameters than observations, as $d = 350 \times 3 = 1050 > n$. Each simulated covariate vector $\mathbf{X}_i$ was a centered and unit-scaled multivariate normal, with AR(0.25) serial collinearity. To assess the performance of the fusion penalty, we lastly varied the extent of shared covariate effects. Under the ‘Shared Support’ specification, the support of the non-zero effects is the same across submodels, whereas under the ‘Partially Non-Overlapping Support’ structure the supports only partially overlap.

5.2 Analyses

Under each scenario, we generated 300 simulated datasets. Each dataset was then analyzed using both Lasso and SCAD-penalized models, each with and without additional fusion ℓ_1-penalties linking all three hazards. Each analysis was performed using both Weibull and piecewise constant baseline hazard specifications. For the latter, we set $k_g = 3$ and chose breakpoints at quantiles of the data, so they also did not overlap exactly with the true data generating mechanism. In all cases, penalized models were fit over a grid comprising 21 values for λ_1 in the high-dimensional setting and 29 in the low-dimensional setting, and 4 values for λ_2, leading to overall regularization grids of 21×4 = 84 and 29×4 = 116 points, respectively. At each grid point, the best estimate was selected from initializations at the previous step’s solution, and 5 additional randomized starting values.

From each fitted regularization grid, two models are reported. A model without fusion was chosen that minimizes the BIC over the subset path $(\lambda_1, 0)$, and a model possibly with fusion was chosen which minimized BIC over the entire grid of values (λ_1, λ_2). Therefore, the model space with fusion encompasses the model space without fusion, so any differences between the reported estimates with and without fusion reflect improvements in the BIC due to the added fusion penalty. If fusion did not improve BIC, there would be no difference.

For comparison, we considered a forward selection procedure minimizing BIC by adding one covariate to one transition hazard at each step. Finally, we fit the 'oracle' MLE on the set of true non-zero coefficients, as well as the full MLE in the low-dimensional setting.

5.3 Results

To assess estimation performance, we examine ℓ_2-error defined as $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ in Table 1. Across all settings, the estimation error of the regression coefficients was insensitive to the model baseline hazard specification, with comparable results for both Weibull and piecewise constant specifications. Therefore we present results for Weibull models, with piecewise specification results given in Web Appendix F. For both $n = 500$ and $n = 1000$, the combination of SCAD and fusion penalties outperforms all comparators, particularly in the high-dimensional regime. Forward selection and the unfused SCAD-penalized estimator generally yielded the next best results. Estimators with fusion penalization also performed better in the 'Low Non-Terminal Event Rate' setting, likely because fusion links estimates across submodels, allowing 'borrowing' of information about h_1 and h_3 from h_2 when the non-terminal event is rare. Fusion penalized estimators also performed comparably well even if the covariate supports of each submodel only partially overlapped, relative to complete overlap. However, Lasso penalized models did poorly relative to other comparators, which likely reflects elevated regularization-induced bias in the individual estimates.

To assess selection performance, Table 2 reports mean sign inconsistency, which counts the estimated regression coefficients that do not have the correct sign—exclusion of true non-zero coefficients, inclusion of true zero coefficients, or estimates having the opposite sign of the true coefficient. Lower values indicate better overall model selection performance. Again performance was very similar between Weibull and piecewise constant specifications, so only Weibull results are presented in the main text. Additional simulation results, and separated results on false inclusions and exclusions are included in Web Appendix F.

For both $n = 500$ and $n = 1000$, the combination of SCAD and fusion penalties out-performed comparators, while other methods' performances varied across sample size and setting. Fusion penalized estimators exhibited notably better selection properties in the 'Shared Support' setting, as fusion coerces a common block of non-zero covariates across submodels. Lasso penalized models tended to choose overly sparse models. With many of the true non-zero effects small in magnitude, regularization-induced bias may have rendered those terms indistinguishable from truly zero effects.

Lastly we summarize the results in the smallest sample setting of $n = 300$, which are detailed in Web Appendix F. This is a challenging setting because small samples exacerbate the non-convexity of the marginal illness-death likelihood, and when outcomes are rare some transition submodels have a small number of observed events. These complications affect small-sample empirical performance of frailty-based illness-death models even in the absence of high-dimensional covariates. For example, in the setting with 25 covariates per transition, average estimation error of the full MLE is substantially larger for $n = 300$ than $n = 500$, and more sensitive to the non-terminal event rate (see Web Table F.4). Still, in this low-dimensional regime we again observe the combination of SCAD and fusion penalties reducing estimation error relative to comparators.

However, increasing dimensionality to 350 covariates per transition while keeping $n =$ 300 degraded performance of all methods, particularly when event rates were lowest. For example, in the ‘Low Non-Terminal Event Rate’ settings, the comparator forward selection algorithm failed for between 25 and 90 percent of simulations by adding so many covariates that optimization no longer converged. Penalized models also tended towards extremes, with SCAD-penalized models including many unnecessary covariate effects, while the Lasso models selecting few or no non-zero covariate effects. A key challenge is that likelihood-based selection criteria like BIC can be distorted in small samples by non-convexity. In certain instances, the log-likelihood can become monotonic with respect to the frailty log-variance σ, yielding artificial information criteria and leading to selected models that are either completely sparse (as with Lasso-penalized models) or completely saturated (as with SCAD-penalized and forward-selected models). These serious complications manifested only in the most difficult settings combining small samples, low-to-moderate event rates, and high dimensional covariates, but show that the challenges of very small-sample estimation with frailties is compounded by high-dimensional covariates.

6. Data Application: Preeclampsia (PE) and Delivery

The proposed methods are motivated by practical application to clinical settings where interest is in developing a risk model that jointly characterizes a non-terminal and terminal event. To this end, we consider modeling PE onset and the timing of delivery using the electronic health records of an urban, academic medical center in Boston, Massachusetts. We analyze 2127 singleton live births recorded in 2019 among individuals without pre-existing hypertension who received the majority of their prenatal care and delivered at the academic medical center. Restricting to those without hypertension targets the modeling task, as PE superimposed on chronic hypertension has distinct clinical features compared to other forms of the disease (Jeyabalan, 2013). 189 (8.9%) individuals developed PE, with median diagnosis time of 37.9 weeks (Inter-Quartile Range [IQR] 35.0–39.0). The median time to delivery was 38.0 weeks (IQR 35.4–39.3) among those who developed PE and 39.4 weeks (IQR 38.6–40.3) among those who did not. Note, because PE is only diagnosed after 20 weeks of gestation, for modeling purposes this is used as the time origin, with T_1 and T_2 defined as time from week 20 until PE onset and delivery, respectively.

We considered a set of 33 potential covariates, including demographics recorded at patient intake, baseline lab values annotated by the medical center with a binary indicator for abnormality, and maternal health history derived from ICD-10 diagnostic codes associated with delivery (summarized in Web Table A.1). We fit Markov illness-death models so that $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are both interpretable on the gestational age timescale, under both Weibull and piecewise constant baseline hazards. We adopted SCAD penalties on each regression coefficient, and an ℓ_1 fusion penalty between $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ to induce a shared structure between coefficients for the timing of delivery in the absence of PE and timing of delivery given the onset of PE. We specified a grid of 55 values for λ_1 and seven for λ_2. As above, we selected the final penalized model minimizing BIC over the entire grid of (λ_1, λ_2) values, while the final SCAD-only penalized model minimizes BIC over the path $(\lambda_1, 0)$. Again, this means that differences between the models with and without fusion reflect improvements in BIC due to the fused penalty. For comparison, we also report the unpenalized MLE.

Figure 2 compares the estimated regression coefficients between unpenalized and penalized models, for each baseline hazard specification. As in the simulation studies, there appears to be little difference in the selection properties or resulting regression estimates between models with Weibull and piecewise constant baseline hazard specifications. Across all specifications, the penalized estimates chosen by BIC are highly sparse relative to the unpenalized MLE. The inclusion of a fusion penalty linking the coefficients in h_2 and h_3 further improved BIC, and fused several covariate effects for the timing of delivery before PE and given PE. Specifically, parity of 1 or more (meaning a history of at least one pregnancy lasting at least 20 weeks), and in the Weibull model, the presence of leiomyomas (benign gynecological tumors), are both estimated with shared coefficients on timing of delivery with and without PE. For comparison, models chosen by AIC are provided in Web Figure A.1, which include more covariates but are still sparse relative to the full model. Finally, we find that in this application, beyond the estimated regression coefficients frailties did not play a large role in characterizing additional dependence between preeclampsia and delivery timing, as the estimated frailty variance is very close to 0 in all estimates.

These results have strong clinical significance. In every specification and across all three hazards, the selected covariates are primarily maternal health history and behaviors, rather than demographics or baseline lab measurements. Many of the variables selected for the cause-specific hazard of PE—parity of 1 or more, BMI of at least 30, and pre-existing diabetes—align with findings of recent meta-analyses of factors affecting PE risk (Giannakou et al., 2018). Further illustrating the interplay of risk factors with the outcomes, fusion penalized estimates show parity of 1 or more associated both with delayed timing of PE, and accelerated timing of delivery in the presence of PE. This correspond clinically with risk of milder late-onset PE for which delivery can occur quickly with fewer risks.

As introduced previously, care decisions for PE center two challenges: identifying those at high risk of PE, and timing delivery after PE onset to balance maternal and fetal health risks. Though our methodological focus is on regularized estimation and model selection, the resulting fitted illness-death models also generate prospective risk predictions to inform these individualized clinical care decisions (Putter et al., 2007). In Web Appendix B we present and discuss a set of four such risk profiles for sample patients using a Weibull model with fusion penalty. Specifically, from baseline the model can predict across time how likely an individual is to be in one of four categories: (i) still pregnant without PE, (ii) already delivered without PE, (iii) already delivered with PE, and (iv) still pregnant with PE. Such profiles directly address clinical needs by highlighting individuals' overall risk of developing PE, while also characterizing the timing of PE and delivery.

7. Discussion

Frailty-based illness-death models enable investigation of the complex interplay between baseline covariates and semi-competing time-to-event outcomes. Estimates directly illustrate the relationships between risk factors and the joint outcomes via hazard ratios across three submodels, while individualized risk predictions generate an entire prospective outcome trajectory to inform nuanced clinical care decisions.

The task of modeling risk of PE and timing of delivery illustrates the value, and the potential, for penalized illness-death modeling to inform clinical practice. While analysts typically default to including the same set of covariates in all three hazards, Figure 2 illustrates that no covariate in any BIC-selected models has distinct, non-shared coefficients in all three hazards. Moreover, even in the setting of PE onset and delivery, where relatively few covariate effects appear to be shared across submodels, adding fusion regularization improved model fit metrics. We expect the impact of fusion would be even more pronounced in settings where the outcomes are more positively correlated, such as when the non-terminal event is a negative health outcome and the terminal event is death. Because frailties also tend to characterize positive correlation of the outcomes, we might also expect larger estimated frailty variance in such settings. Analysts interested in considering non-frailty models might fit regularization paths with and without the frailty, and choose a final criterion-minimizing model from amongst both frailty and non-frailty candidates.

We also note that the statistical rate result of Theorem 1 uses the specific choice of penalty given in (10), however we would expect similar theoretical performance under the similar penalty introduced in (5). The advantage of implementing the formulation as in (5) is its interpretability for the analyst, by directly distinguishing between the role of λ_1 in determining the global level of sparsity of the regression parameters, and the role of λ_2 in determining the level of parsimony in the sharing of effects across hazards.

Though the current work focuses on penalization of the regression parameters $\boldsymbol{\beta}$, the framework also admits penalization of the baseline hazard parameters to achieve similar goals of flexibility and structure. For example, under the Markov transition specification a penalty of the form $\sum_{j=1}^{k_3} p_{\lambda_3}(|\phi_{2j} - \phi_{3j}|)$ could regularize the model towards having $h_{02}(t_2) = h_{03}(t_2)$. Xu et al. (2010) call this the ‘restricted’ illness-death model corresponding to the baseline hazard of the terminal event being equal before and after the non-terminal event. Moreover, while we presently focus on fixed-dimensional parametric baseline hazard specifications, in principle the estimation algorithms presented here extend to penalized baseline models of growing dimensionality, such as splines with number of basis functions dependent on sample size. The theoretical properties of such an estimator would be an interesting avenue of future research. Future work might also explore these methods and theory under other frailty distributions besides the closed form-inducing gamma.

Finally, establishing the statistical rate of the proposed penalized estimator also enables future development of post-selection inferential tools such as confidence intervals for selected coefficients. Most importantly, this methodology enables future work modeling semi-competing risks across a wide array of clinical domains, and leveraging data sources with high-dimensional covariates from electronic health records to genomic data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Associate Editor and Reviewers for helpful comments, and thank Drs. Michele Hacker, Anna Modest, and Stefania Papatheodorou for data access at Beth Israel Deaconess Medical Center, Boston, MA and valuable discussion. This study received exemption status by institutional review boards at Beth Israel Deaconess Medical Center and Harvard TH Chan School of Public Health. HTR is supported by NIH grants T32LM012411, T32GM074897, and F31HD102159, and JL by R35CA220523 and U01CA209414. JL is supported by NSF grant 1916211.

Data Availability Statement

Data used in this paper to illustrate the proposed methods are not shared due to privacy restrictions.

References

Chapple AG, Vannucci M, Thall PF, and Lin S (2017). Bayesian variable selection for a semi-competing risks model with three hazard functions. Computational Statistics & Data Analysis 112, 170–185. [PubMed: 29033478]

Chen X, Lin Q, Kim S, Carbonell JG, and Xing EP (2012). Smoothing proximal gradient method for general structured sparse regression. Annals of Applied Statistics 6, 719–752.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fine J, Jiang H, and Chappell R (2001). On semi-competing risks data. Biometrika 88, 907–919.

Giannakou K, Evangelou E, and Papatheodorou SI (2018). Genetic and non-genetic risk factors for pre-eclampsia: Umbrella review of systematic reviews and meta-analyses of observational studies. Ultrasound in Obstetrics & Gynecology 51, 720–730. [PubMed: 29143991]

Jazić I, Schrag D, Sargent DJ, and Haneuse S (2016). Beyond composite endpoints analysis: semicompeting risks as an underutilized framework for cancer research. JNCI: Journal of the National Cancer Institute 108, djw163. [PubMed: 27389914]

Jeyabalan A (2013). Epidemiology of preeclampsia: impact of obesity. Nutrition Reviews 71, S18–S25. [PubMed: 24147919]

Lee KH, Haneuse S, Schrag D, and Dominici F (2015). Bayesian semiparametric analysis of semicompeting risks data: investigating hospital readmission after a pancreatic cancer diagnosis. Journal of the Royal Statistical Society: Series C (Applied Statistics) 64, 253–273. [PubMed: 25977592]

Loh PL and Wainwright MJ (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research 16, 559–616.

Putter H, Fiocco M, and Geskus RB (2007). Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 26, 2389–2430. [PubMed: 17031868]

Sennhenn-Reulen H and Kneib T (2016). Structured fusion lasso penalized multi-state models. Statistics in Medicine 35, 4637–4659. [PubMed: 27334132]

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58, 267–288.

Tibshirani RJ and Taylor J (2011). The solution path of the generalized lasso. Annals of Statistics 39, 1335–1371.

Wang Z, Liu H, and Zhang T (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Annals of Statistics 42, 2164–2201. [PubMed: 25544785]

Xu J, Kalbfleisch JD, and Tai B (2010). Statistical analysis of illness–death processes and semicompeting risks data. Biometrics 66, 716–725. [PubMed: 19912171]

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38, 894–942.

Zhao T, Liu H, and Zhang T (2018). Pathwise coordinate optimization for sparse learning: Algorithm and theory. Annals of Statistics 46, 180–218.

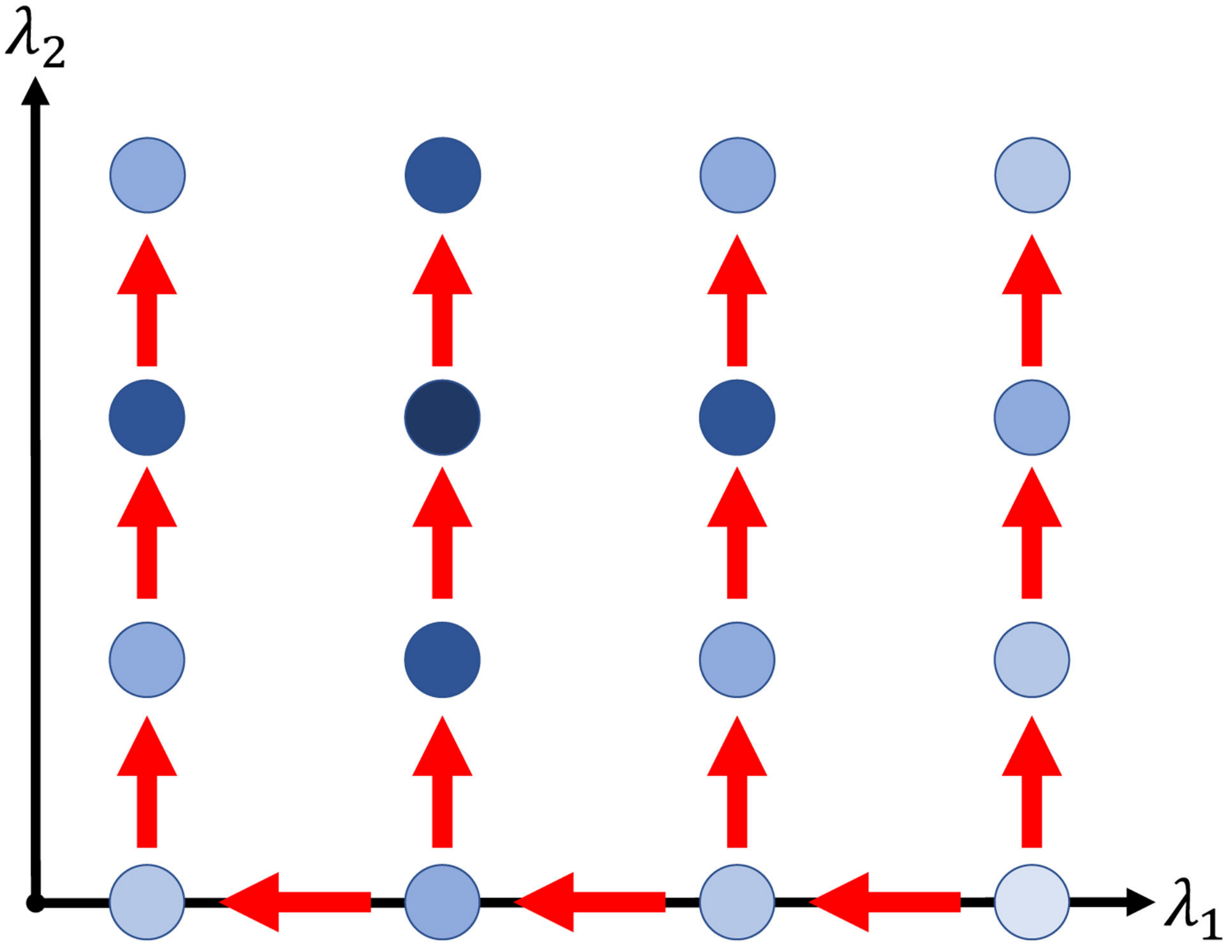

Figure 1.

Schematic depicting path-following grid search routine over (λ_1, λ_2). Each dot represents a (λ_1, λ_2) pair for which the penalized estimator is fitted, with darker shading corresponding to better model fit metric (e.g., AIC or BIC). The arrows illustrate the path of the search routine, with optimization at each grid point starting at the solution of the previous point. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

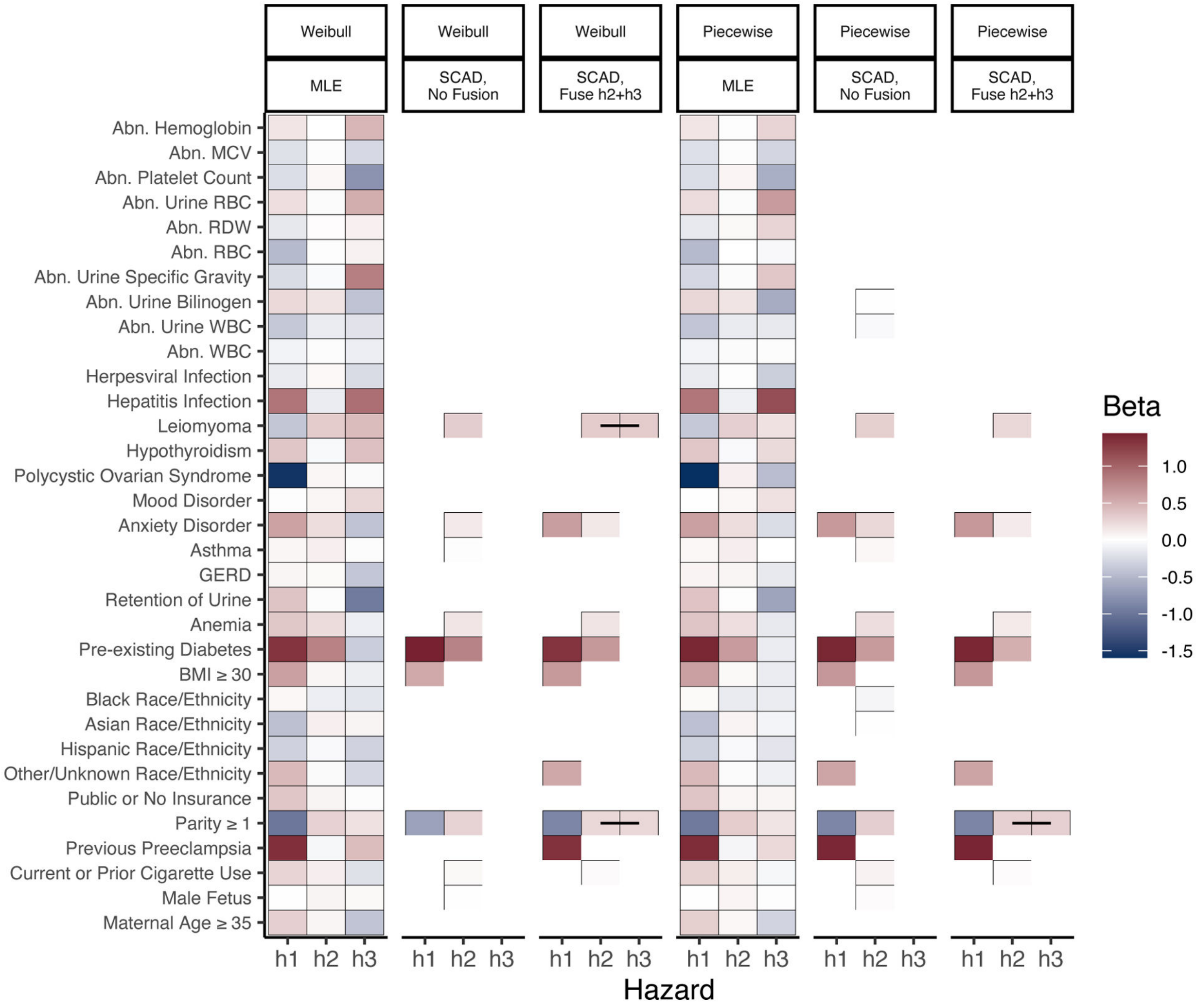

Figure 2.

Estimated coefficients, BIC-optimal SCAD-penalized estimators with and without ℓ_1 fusion between h_2 and h_3, and MLE under Markov specification. Fused coefficients connected with a black line. Abbreviations: abnormal (Abn), white blood cell count (WBC), red blood cell count (RBC), red cell distribution width (RDW), mean corpuscular volume (MCV), gastroesophageal reflux disease (GERD). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1

Mean ℓ_2 estimation error of $\widehat{\beta}$, Weibull baseline hazard specification. Maximum likelihood estimates only available for low-dimensional setting.

***n* = 500**	**Oracle**	**MLE**	**Forward**	**Lasso**	**SCAD**	**Lasso + Fusion**	**SCAD + Fusion**
Moderate Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.75	1.34	1.48	2.06	1.35	1.49	0.97
High-Dimension	0.76	—	3.10	2.77	2.37	2.21	1.20
Partially Non-Overlapping Support							
Low-Dimension	0.73	1.30	1.33	1.87	1.26	1.61	1.13
High-Dimension	0.74	—	3.47	2.49	2.34	2.28	1.45
Low Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.92	1.81	1.89	2.24	1.91	1.48	1.20
High-Dimension	0.88	—	5.13	2.56	2.23	2.34	1.28
Partially Non-Overlapping Support							
Low-Dimension	0.80	1.53	1.55	2.05	1.50	1.71	1.27
High-Dimension	0.80	—	3.65	2.42	2.20	2.33	1.55

***n* = 1000**	**Oracle**	**MLE**	**Forward**	**Lasso**	**SCAD**	**Lasso + Fusion**	**SCAD + Fusion**
Moderate Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.50	0.82	0.76	1.53	0.73	1.39	0.75
High-Dimension	0.50	—	1.25	2.40	1.52	1.82	0.81
Partially Non-Overlapping Support							
Low-Dimension	0.48	0.81	0.71	1.23	0.71	1.32	0.83
High-Dimension	0.49	—	1.18	2.17	1.20	1.84	0.97
Low Non-Terminal Event Rate							
Shared Support							
Low-Dimension	0.57	0.95	1.31	2.05	1.15	1.37	0.83
High-Dimension	0.58	—	1.86	2.39	2.10	1.80	0.90
Partially Non-Overlapping Support							
Low-Dimension	0.52	0.88	0.85	1.52	0.84	1.43	0.96
High-Dimension	0.52	—	1.42	2.22	1.80	1.94	1.06

Table 2

Mean count of sign-inconsistent $\widehat{\boldsymbol{\beta}}$ estimates, Weibull baseline hazard specification. Sign inconsistency counts the number of estimated regression coefficients that do not have the correct sign—exclusion of true non-zero coefficients, inclusion of true zero coefficients, or estimates having the opposite sign of the true coefficient.

***n* = 500**	**Oracle**	**Forward**	**Lasso**	**SCAD**	**Lasso + Fusion**	**SCAD + Fusion**
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.13	11.51	15.06	10.28	11.90	3.45
High-Dimension	0.12	35.81	26.40	35.22	21.89	18.88
Partially Non-Overlapping Support						
Low-Dimension	0.14	10.73	15.39	10.19	15.06	7.91
High-Dimension	0.13	34.56	27.21	38.52	23.37	24.35
Low Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.30	16.23	19.33	17.16	12.77	5.45
High-Dimension	0.29	39.50	26.08	24.94	23.39	14.42
Partially Non-Overlapping Support						
Low-Dimension	0.24	12.70	17.11	12.42	16.33	9.47
High-Dimension	0.21	36.20	26.08	29.51	24.53	22.89
***n* = 1000**	**Oracle**	**Forward**	**Lasso**	**SCAD**	**Lasso + Fusion**	**SCAD + Fusion**
Moderate Non-Terminal Event Rate						
Shared Support						
Low-Dimension	0.01	4.03	13.52	3.99	8.48	1.51
High-Dimension	0.00	15.83	20.92	19.92	19.82	7.21
Partially Non-Overlapping Support						
Low-Dimension	0.02	4.12	13.80	4.27	12.58	4.00
High-Dimension	0.03	15.57	19.86	19.30	21.28	8.44
Low Non-TerminaEvent Rate						
Shared Support						
Low-Dimension	0.06	9.70	16.60	9.01	8.43	1.86
High-Dimension	0.06	23.73	24.30	22.17	19.08	6.95
Partially Non-Overlapping Support						
Low-Dimension	0.03	5.29	14.54	5.74	13.27	4.97
High-Dimension	0.03	16.96	21.98	22.18	21.98	8.72