Flexible Competing Risk Modeling for Big Data from Administrative Records and Disease Registries

by

Wenbo Wu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics and Scientific Computing)
in The University of Michigan
2022

Doctoral Committee:

Research Associate Professor Kevin He, Chair
Professor Emeritus John D. Kalbfleisch
Professor Jian Kang
Professor Joseph M. Messana
Professor Jeremy M.G. Taylor

Wenbo Wu

wenbowu@umich.edu

ORCID iD: 0000-0002-7642-9773

To Yiting, Qiulan, and Shouyi

ACKNOWLEDGEMENTS

Reflecting on my predoctoral adventure at Michigan, I am deeply indebted to so many people without whom this dissertation would have never come to completion.

I must start by thanking my advisor and dissertation committee chair Kevin He, whose transcendent mentorship has gotten me through uncountable bogs in research with sustained optimism and progress. I am extremely fortunate to have worked closely with Jack Kalfbleisch. His unsurpassed expertise and enthusiasm have profoundly shaped my perspective in statistics both as a subject and as a career. I want to thank my dissertation committee members Jeremy Taylor, Jian Kang, and Joe Messana. Jeremy and Jian have been persistently approachable whenever I need their insights. Joe has been incredible as a nephrologist with a solid statistical understanding. Our collaborations at the Kidney Epidemiology and Cost Center have always been fruitful. I would also like to extend my sincere gratitude to Kirsten Herold. She has demonstrated herself as an exceptional educator with ample experience in scientific writing and career development.

I also owe a debt of thanks to Valarie Ashby, Lonnie Barnes, Andrew Brouwer, Richard Burney, Shu Chen, Claudia Dahlerus, Xuemei Ding, Ashley Eckard, Wolf Gremel (Optum, Inc.), Justin Hagar, Nick Hartman, Rich Hirth, Hui Jiang, Jonathan Kuriakose (Rutgers University), Rahul Ladhania, John Magee, Elham Mahmoudi, Jeet Naik, Robin Padilla, Katrina Price, Doug Schaubel (University of Pennsylvania), Rajiv Saran, Jon Segal, Ananda Sen, Vahakn Shahinian, Tempie Shearon, Xu

Shi, Megan Slowey, Randy Sung, Anca Tilea, Mia Wang, Wenjing Weng, Lu Xia (University of Washington), Tao Xu, and Xiaosong Zhang.

I would be remiss not to thank Yiting Li for her relentless support and encouragement as my wife. Loving an academic is no small chore and she has been there for my late nights, deadlines, rejections, and of course, accomplishments. This dissertation is also dedicated to my parents and the memory of my grandparents.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Competing risks are omnipresent in administrative records and disease registries. The increasing availability of data facilitates a comprehensive investigation on competing risks in various contexts, potentially leading to significant improvements in health care quality, and a deeper understanding of the etiology of deadly diseases. At the same time, the growing volume of data, high-dimensional parameter space, and complexity of modeling necessitate methodological advances beyond existing analytical frameworks. In this dissertation, we develop novel statistical and computational methods for profiling health care providers and characterizing the variation of coefficients of risk factors. These methods are specifically tailored to large-scale competing risks data.

The 30-day hospital readmission rate has been widely used in profiling hospitals and dialysis facilities, among other health care providers. Current analyses typically use logistic regression to model readmission as a binary outcome without explicitly considering competing risks (e.g., death). This oversight leads to less comprehensive modeling and distorted provider evaluation. To address these drawbacks, we propose a discrete-time competing risk model, where the cause-specific readmission hazard is used to assess provider-level effects. This readmission-focused assessment utilizes the standardized readmission ratio as the associated quality measure; this ratio is not systematically affected by the rate of competing risks. To facilitate the estimation and inference of thousands of provider effects, we develop

an efficient Blockwise Inversion Newton algorithm, and a stabilized robust score test that overcomes the conservative nature of the classical robust score test. An application to Medicare dialysis patients demonstrates improved profiling, model fitting, and outlier detection over existing methods.

Time-varying coefficient modeling has proven useful for competing risk analysis. When examining the cause-specific etiology of breast and prostate cancers using the large-scale data from the Surveillance, Epidemiology, and End Results (SEER) Program, we encountered two challenges that existing time-varying coefficient models cannot tackle. First, these methods, dependent on expanding the original observations as repeated measurements, result in formidable time and memory consumption as the sample size escalates to over one million. Second, when binary predictors are present with near-zero variance, existing methods suffer from numerical instability and inaccurate estimation due to ill-conditioned second-order information. To address these issues, we propose a proximal Newton algorithm with a shared-memory parallelization scheme. Applications to the SEER data demonstrate that effects of tumor stage on cause-specific deaths vary substantially with the time since diagnosis.

Our investigation into the impact of COVID-19 on dialysis patients suggests that effects of COVID-19 on post-discharge outcomes vary with both post-discharge and calendar time. This evidence motivates us to develop a novel varying coefficient model, where each coefficient is a bivariate function of the event time and an external covariate. The model leverages tensor-product B-splines to account for the coefficient variation in two dimensions. Difference-based anisotropic penalization is introduced to mitigate model overfitting and the wiggliness of the estimated trajectories; various cross-validation methods are considered in the determination of optimal tuning parameters. Hypothesis testing procedures are designed to examine whether

the COVID-19 effect varies significantly with post-discharge time and the time since pandemic onset. Simulation experiments are conducted to evaluate the estimation accuracy, type I error rate, statistical power, and model selection procedures. Applications to Medicare dialysis patients demonstrate the real-world performance of the proposed methods.

Overall, the approaches presented here offer promising avenues for analyzing high-volume competing risks data of multilevel and multidimensional structure.

CHAPTER I

Introduction

Terminal events of distinct failure types, called competing risks [59], are commonly encountered in administrative records (e.g., Medicare) and disease registries (e.g., the Surveillance, Epidemiology, and End Results Program, SEER). Such databases typically contain a variety of patient characteristics (demographic, procedural and clinical) from thousands of or even millions of individuals [15, 84]. For instance, 1,093,192 female patients first diagnosed with breast cancer between 1973 and 2015 are included in the SEER breast cancer sample, along with their age, race, tumor stage, and their dates and causes of death, if not censored; in the 2018 sample of hospital discharges among Medicare dialysis patients, there are 272,897 patients in 6,937 Medicare-certified dialysis facilities with 541,769 qualifying discharges, along with patients' demographics, clinical characteristics, prevalent comorbidities, and post-discharge event times.

With the wealth of information in these large-scale databases, we are in a good position to advance our understanding of critical issues in population health, which potentially informs health care policy and improves clinical practice. As described above, the Medicare claims database contains information about post-discharge outcomes of patients with end-stage renal disease (ESRD), who are associated with all

Medicare-certified dialysis facilities in the United States. This unique source can be leveraged toward a payment-linked assessment of dialysis facilities based on post-discharge outcomes, thereby promoting high-quality health services among dialysis facilities [16]. Alternatively, the Medicare claims data for dialysis patients can be used to examine the possibly varying effect of a particular risk factor, e.g., the coronavirus disease 2019 (COVID-19). As a second example, the cause-specific death data from the SEER registry can elucidate the etiology of a type of cancer as well as the disease progression over the post-diagnosis period. Any dynamic effects of risk factors on cancer survival may have implications for risk prediction, treatment, and care [11].

Despite the improved statistical power and external validity, competing risks data of immense volume and complex structure pose statistical and computational challenges that existing methods cannot properly address. For example, the non-trivial presence of competing risks renders existing methods conceptually flawed for readmission-focused provider profiling. Specifically, these methods, typically treating readmission as a binary outcome [49, 37, 36, 87, 88, 3, 80], are highly sensitive to the rate of competing risks, and hence result in distorted provider evaluation. The accompanying software implementations [99], developed for moderate-sized and general-purpose data analytics, impose unwieldy time and memory burdens to the simultaneous estimation and (or) inference of thousands of unknown model parameters. In addition, when the interest is focused on the cause-specific varying effects of risk factors, existing methods also suffer from numerical instability and inaccurate estimation especially in the context of ill-conditioned second-order information.

As analytical strategies targeted at the aforementioned challenges, in this dissertation, we develop flexible statistical and computational methods for large-scale

competing risks data from administrative records and disease registries. In Chapter II, we propose a discrete-time competing risk model, wherein the cause-specific readmission hazard is used for assessing health care providers. In Chapter III, we consider a proximal Newton algorithm with a shared-memory parallelization scheme and a test of nonproportionality for time-varying effects. Chapter IV introduces a bivariate varying coefficient model featuring tensor-product B-splines [106] and difference-based anisotropic penalization [132].

CHAPTER II

Analysis of Hospital Readmissions with Competing Risks

2.1 Introduction

In 2019, approximately 37 million American adults (15%) were estimated to have chronic kidney disease (CKD), the ninth leading cause of death in the United States [14]. Dialysis patients with end-stage renal disease (ESRD), the final stage of CKD, experience high mortality rates [25], and frequent hospital admissions and readmissions [104]. On average, ESRD patients are hospitalized almost twice a year, with more than one in three hospital discharges followed by readmission within 30 days. According to recent estimates, treatment of ESRD patients costs 7.2% of total Medicare expenditures [104].

To improve quality of care and reduce costs for dialysis patients, the Centers for Medicare and Medicaid Services (CMS) monitors Medicare-certified dialysis facilities nationwide with various quality measures of patient outcomes (e.g., mortality and hospitalization) and provides feedback and information to facilities, patients and other stakeholders. One outcome of particular interest is unplanned hospital readmission (UHR). Clinical evidence supports UHR as an indicator of facility-hospital care coordination, medical cost-effectiveness, and patient quality of life [17, 13, 100, 127]. The facility-level 30-day UHR is typically measured by the standardized readmis-

sion ratio (SRR), which is the ratio of the number of observed UHRs to the number expected with reference to a national average facility (a national norm) given the patient characteristics of that facility. An SRR less (or greater) than one implies that the facility's observed readmission rate is lower (or higher) than expected based on the national norm. The SRR has been implemented by CMS in its ESRD Quality Incentive Program, a value-based purchasing program, in which payment for treating patients is linked to a facility's quality measures [16]. Thus, valid assessment of dialysis facilities according to UHR has important implications for health care quality evaluation and policy making.

In this chapter we develop statistical methods for readmission-focused provider profiling. Current modeling frameworks treat UHR as a binary outcome and use logistic regression with fixed or random facility effects [49, 37, 36, 87, 88, 3, 80]. These analyses are routinely used by CMS in calculating readmission measures for hospital or dialysis facility evaluation [53, 62].

In practice, however, a patient may experience a competing event, such as death, prior to a UHR during the follow-up. In our motivating example, 15.31% of discharges with subsequent events within 30 days were initially followed by competing events, which were distributed among 6,230 (89.8%) facilities. The logistic model only considers the occurrence of an event during follow-up, regardless of the time at which the event occurs. As a consequence, an early UHR on Day 4 is deemed equivalent to a late UHR on Day 29 even though the latter is very close to the end of the period of interest. Similarly, one competing event on Day 4 and another on Day 29 are deemed equivalent, although in the first case, the individual is only at risk of a UHR for 4 days whereas the latter is at risk for 29 days.

We propose a discrete time competing risk model based on a cause-specific

hazards framework [98], in which UHRs constitute one failure type and competing events (e.g. death, planned readmission, etc.) constitute the other. This model then accounts for the timing of events and distinguishes between an early and late event, which offer several advantages over existing approaches to modeling UHR. Based on the competing risk model, we find that current methods essentially consider a logistic model for the cumulative incidence function of a UHR evaluated at 30 days; this approach makes no distinction between early and late events, and does not take into account the number of competing events. An unintended consequence is that a given facility may appear to have a lower readmission rate due to having a higher rate of competing risks.

Applying the competing risk model to the CMS readmissions data poses challenges to estimation and inference. First, fitting the proposed model via the maximum likelihood approach involves estimating 6,937 facility-specific fixed effects along with coefficients of risk factors, which poses considerable computational challenges. To address the computational issues, we develop a Blockwise Inversion Newton (BIN) algorithm applying the block matrix inversion formula to the associated Fisher information matrix. Compared with existing algorithms, the fast-converging BIN achieves scalability and memory efficiency.

Second, the 541,769 discharges in our data were associated with 272,897 patients from 6,937 facilities, of which 727 (10.48%) had fewer than 25 discharges and 704 (10.15%) had readmission rates of 15.4% or lower. The presence of patient-level clustering and small facilities with low readmission rates renders existing inferential methods ill-suited for testing the facility effects. To name a few, the exact test by He et al. [49] overlooks patient-level correlation; robust or not, the numerically unstable Wald tests often yield inflated type I errors [49, 124]; and the robust estimator

by Liang and Zeger [75] deteriorates as cluster size shrinks. Inspired by Pan [90] and Rotnitzky and Powell [102], we devise a novel robust score test, which stabilizes the variance estimation by integrating correlation information across patients from a large number of facilities. Compared with alternative approaches, this test demonstrates enhanced power, controlled type I error, and less skewed outlier detection.

The rest of this chapter is organized as follows: Section 2.2 develops the competing risk model.Section 2.3 proposes techniques for estimation and inference. In Section 2.4, we apply the proposed methods to 30-day readmissions among ESRD patients, using data obtained from CMS. Section 2.5 presents simulation results and Section 2.6 concludes with a discussion.

2.2 Discrete Competing Risk Model

2.2.1 The Model

Let the variate T be the time in days until the first post-discharge event and let τ denote the follow-up of interest (e.g., 30 days). Let D indicate whether the event is a UHR ($D = 1$) or a competing risk ($D = 2$). If there is no event up to time τ, then $D = 0$ with $T \geq \tau + 1$. The cause-specific hazard function for event type d is

$$\lambda_d(t) \coloneqq \mathrm{P}(T = t, D = d | T \geq t), \quad t = 1, ..., \tau;\ d = 1, 2$$

where hazard $\lambda_1(t)$ of UHR is of particular interest. It is sometimes said that this formulation assumes that the failure types are independent, but this is not true. They are, however, mutually exclusive and the total hazard at time t is $\lambda_1(t) + \lambda_2(t)$. By the usual results for discrete survival data, the survivor function is

$$S(t) \coloneqq \mathrm{P}(T \geq t) = \prod_{k=0}^{t-1} \{1 - \lambda_1(k) - \lambda_2(k)\}, \quad t = 1, \ldots, \tau,$$

where we assume $\lambda_1(0) = \lambda_2(0) = 0$.

In the continuous cause-specific hazard model, the survivor function is factorizable, but this is not the case in a discrete competing risk model [59, 73]. However, we can make a change from $\lambda_2(t)$ to

$$\mu(t) \coloneqq \mathrm{P}(T = t, D = 2 | T \geq t, D \neq 1) = \lambda_2(t)/\{1 - \lambda_1(t)\}, \quad t = 1, \ldots, \tau,$$

which implies that $1 - \lambda_1(t) - \lambda_2(t) = \{1 - \lambda_1(t)\}\{1 - \mu(t)\}$. As we will see later, this factorization leads to likelihood decomposition which allows modeling the UHRs alone. A symmetric reparameterization applies to $\lambda_2(t)$, although it is of less interest in our application. To simplify the notation, we will hereafter use $\lambda(t)$ to denote the hazard function of UHR. Thus, the survivor function can be rewritten as

$$S(t) = \prod_{k=0}^{t-1} \{1 - \mu(k)\}\{1 - \lambda(k)\}, \quad t = 1, \ldots, \tau.$$

In order to incorporate risk adjustment, it is convenient to introduce counting process notation. For $d = 1, 2$ and $t = 1, \ldots, \tau$, let $N^d(t) \coloneqq I(T \leq t, D = d)$ denote the number of type d events up to time t with $N^d(0) = 0$, where $I(\cdot)$ is an indicator function; let $\Delta N^d(t) \coloneqq N^d(t) - N^d(t-1)$ be an indicator of a type d event at time t; and let $Y(t) \coloneqq I\{\min(T, \tau) \geq t\}$ indicate whether a patient is at risk for an event (UHR or competing risk) at time t.

Let n_i denote the number of discharges within the ith facility ($i = 1, \ldots, m$), where m is the total number of facilities. For discharge j ($j = 1, \ldots, n_i$) from facility i, the observed data are $\mathcal{H}_{ij}(\tau)$, where $\mathcal{H}_{ij}(t) \coloneqq \{(Y_{ij}(k), \Delta N^1_{ij}(k), \Delta N^2_{ij}(k), \mathbf{Z}_{ij}(k))\}_{k=1}^{t}$. In this sequence, $\mathbf{Z}_{ij}(k)$ is a $r \times 1$ vector of covariates, some elements of which may be external time-dependent covariates [59]. Ignoring the patient-level dependence, the pseudo-likelihood is given by

$$\tilde{L} = L \cdot \prod_{i=1}^{m} \prod_{j=1}^{n_i} \prod_{k=1}^{\tau} \left[\mu_{ij}(k)^{\Delta N^2_{ij}(k)} \{1 - \mu_{ij}(k)\}^{1 - \Delta N^2_{ij}(k) - \Delta N^1_{ij}(k)} \right]^{Y_{ij}(k)}, \tag{2.1}$$

in which

$$L = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \prod_{k=1}^{\tau} \left[\lambda_{ij}(k)^{\Delta N_{ij}^1(k)} \{1 - \lambda_{ij}(k)\}^{1-\Delta N_{ij}^1(k)} \right]^{Y_{ij}(k)}$$

is a partial pseudo-likelihood based on $\{\Delta N_{ij}^1(k)\}_{k=1}^{\tau}$ in the sequence $\mathcal{H}_{ij}(\tau)$ [21, 129]. This partial likelihood argument allows us to focus solely on the part of $\tilde{L}$ related to UHRs and safely ignore the nuisance competing risk process without losing important information on UHRs. The intuition behind L is straightforward: at each at-risk time k, a Bernoulli trial is conducted to identify whether a UHR follows discharge j at facility i, regardless of any competing risks. Information on UHRs is largely represented by the product L of individual Bernoulli likelihoods. The omission of patient-level dependence, i.e., multiple discharges from a single patient, still leads to unbiased estimation, but the estimated variance and standard error can be invalid especially when the intra-patient correlation is high. This issue is addressed in Section 2.3.2 via the proposed stabilized robust score test.

We consider a general formulation of the hazard $\lambda_{ij}(k)$ of UHR for time $k = 1, \ldots, \tau$, discharge j and facility i:

$$\lambda_{ij}(k) = h(\eta_k + \gamma_i + \mathbf{Z}_{ij}^{\top}(k)\boldsymbol{\beta}), \tag{2.2}$$

where h denotes a function whose inverse $g : [0, 1] \to [-\infty, \infty]$ is a monotonically increasing and twice differentiable link function with $g(0) = -\infty$, η_k denotes the g-transformed baseline hazard of UHR at time k, γ_i is a fixed effect for facility i, and $\boldsymbol{\beta}$ denotes a coefficient vector associated with $\mathbf{Z}_{ij}(k)$. Letting $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_\tau)^{\top}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^{\top}$, we have the log-partial likelihood

$$\begin{aligned} \ell(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{\tau} Y_{ij}(k) \big[& \Delta N_{ij}^1(k) \log\{h(\eta_k + \gamma_i + \mathbf{Z}_{ij}^{\top}(k)\boldsymbol{\beta})\} \\ & + \{1 - \Delta N_{ij}^1(k)\} \log\{1 - h(\eta_k + \gamma_i + \mathbf{Z}_{ij}^{\top}(k)\boldsymbol{\beta})\} \big], \end{aligned} \tag{2.3}$$

whose score and information matrix are available in Appendix A.1. Common choices of the link function g include complementary log-log (cloglog), log, and logit. The first of these leads to the grouped relative risk model given time-invariant covariates [58], the second is the discrete relative risk model [97], and the third is the discrete logistic model [20]. In our application, since the discrete (daily) hazard $\lambda_{ij}(k)$ is relatively small, all of these links yield similar results (see Figure A.3 of Appendix A.9).

The estimation of $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ (details in Section 2.3) relies on the maximization of (2.3), where for identifiability of parameters, we impose the constraint $\gamma_{\mathrm{M}} := \mathrm{median}(\boldsymbol{\gamma}) = 0$, and assume that γ_{M} represents the national norm [49]. Given the estimates, we can compute the standardized readmission ratio (SRR). Specifically, the SRR of a facility is defined as the ratio of the observed number of UHRs to the number expected with respect to a national norm, adjusting for patient characteristics. For facility i, we have

$$\mathrm{SRR}_i := \frac{O_i}{E_i} = \frac{\sum_{j=1}^{n_i}\sum_{k=1}^{\tau}\Delta N_{ij}^1(k)}{\sum_{j=1}^{n_i}\sum_{k=1}^{\tau}Y_{ij}(k)h(\hat{\eta}_k+\gamma_{\mathrm{M}}+\mathbf{Z}_{ij}^\top(k)\hat{\boldsymbol{\beta}})}, \tag{2.4}$$

where γ_{M} is the national norm, O_i is the UHR count, E_i is the "expected" number of UHRs, obtained as a sum of the conditional expected number on each day given the at-risk information. An SRR less (or greater) than one means that the facility's observed readmission rate is lower (or higher) than expected based on the national norm.

2.2.2 Significance of the Competing Risk Model (CRM)

In the generalized linear models (GLMs) for binary outcomes [49, 37, 36, 87, 88, 3], the event being analyzed is the occurrence of a UHR prior to time τ and prior to any competing risk. In the homogeneous case without risk adjustment, the

probability of this event is

$$\nu = \sum_{t=1}^{\tau} \left[\lambda(t) \prod_{k=0}^{t-1} \{1 - \mu(k)\}\{1 - \lambda(k)\} \right]. \tag{2.5}$$

This probability ν is the cumulative incidence function (subdistribution) of the UHR evaluated at τ, which depends on the entire series of hazards of UHR and competing risks. Then the likelihood contribution for each discharge under the GLM is

$$\nu^{N^1(\tau)}(1-\nu)^{1-N^1(\tau)} = \left(\frac{\nu}{1-\nu}\right)^{N^1(\tau)} (1-\nu). \tag{2.6}$$

Note that (2.6) does not depend on when the UHR is observed, but only whether or not it is observed (i.e., UHR indicator $N^1(\tau)$ at time τ). Similarly, the timing of the competing risks is irrelevant. Unlike the GLM, the CRM has its discharge-specific contribution to L in (2.1) dependent on the at-risk time $\min(T, \tau)$ but not on the hazard series of competing risks. As a consequence, discharges with unequal at-risk times can be distinguished, and the associated SRR is not systematically affected by the rate of competing risks.

To illustrate the influence of competing risks on SRR-based facility assessment, consider an example in which the hazards of UHR and competing risks are constant over time, and the rates of UHRs and competing risks for a national average facility are 0.3 and 0.1 per month (30 days), respectively. Here the focus is on a very large facility (ignoring sampling errors) with a UHR rate of 0.4 and a rate of competing risks of x. The CRM would give an approximate SRR of $0.4/0.3 = 1.33$ for this facility, while the GLM would give an SRR that depends on x. According to (2.5), the probability of observing a UHR before Day 30 for the facility of interest is

$$\frac{0.4[1 - \exp\{-(0.4 + x)\}]}{0.4 + x}.$$

Similarly, the probability of observing a UHR before Day 30 for a national average

facility is $0.3\{1-\exp(-0.4)\}/0.4$. Thus, the GLM-based SRR is approximately

$$\frac{0.16[1-\exp\{-(0.4+x)\}]}{0.3\{1-\exp(-0.4)\}(0.4+x)},$$

a decreasing function of x. In other words, a higher rate of competing risks leads to an overall decrease in the GLM-based SRR. In particular, for $x = 0$, 0.05, and 0.1, the GLM-based SRR equals 1.33, 1.30, and 1.27, respectively. This inverse relationship between GLM-based SRR and the rate of competing risks indicates that the GRM might give a falsely favorable assessment to a facility with a high rate of UHRs, simply because the facility also has a high rate of competing risks. On the other hand, since the CRM-based SRR remains nearly constant as the rate x of competing risks varies, the CRM would assess two facilities with the same UHR rate as similar, regardless of their difference in the rate of competing risks. This seems a preferable conclusion.

2.3 Estimation and Inference

2.3.1 Blockwise Inversion Newton Algorithm

As noted before, fitting the CRM to our motivating data poses a computational challenge that existing methods cannot tackle. Motivated by Prentice and Gloeckler [96] and Lin and Zhu [76], we develop a Blockwise Inversion Newton (BIN) algorithm (details available in Appendix A.2). This algorithm features efficient inversion of the information matrix of (2.3) via the blockwise inversion formula [9], exploiting the diagonal information submatrix for facility effects.

An analysis of time complexity reveals that the inversion of the information matrix at each iteration of the BIN algorithm costs $O(m(\tau+r)^2+(\tau+r)^3)$, much less than $O((m+\tau+r)^3)$ using a naive Newton–Raphson algorithm given that $m \gg \tau + r$. Because of this substantial efficiency gain, the BIN outperforms existing

software packages such as discSurv [126], which relies on expanding the original data by at-risk time. This advantage is illustrated in Figure 2.1 using real data, which compares the runtime and input data size of BIN and discSurv for varying numbers of facilities.

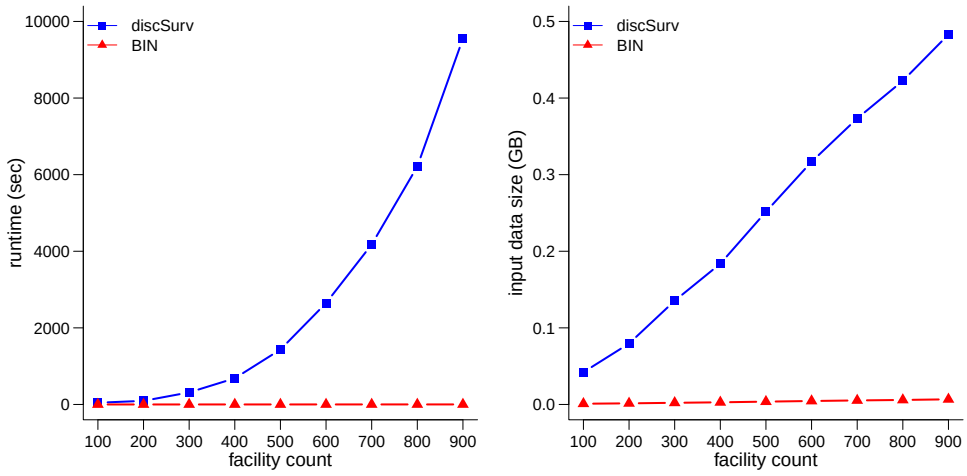


Figure 2.1: Runtime (left) and input data size (right) of BIN and discSurv (data expansion by at-risk time) with varying facility counts. Experiments conducted on an Intel® Xeon® Gold 6254 quad-processor with max frequency 4GHz and RAM 576GB. BIN is implemented using Rcpp [33, 32] and RcppArmadillo [34]. Only two covariates are included for simplicity. discSurv with a facility count beyond 900 induces frequent system freezes as a consequence of data expansion (with over 2 million data records). In contrast, fitting the CRM using BIN to the full-fledged readmissions data (0.335GB) takes 39.512 seconds with 6,937 facilities and 74 covariates.

2.3.2 Stabilized Robust Score Test

When profiling facilities with the SRR, we are interested in the extent to which a facility's SRR differs from one. This suggests testing the null hypothesis $H_{0i} : \gamma_i = \gamma_M$. To incorporate repeated events and small facilities with low readmission rates, we propose a novel robust score test motivated by Pan (2001) [90] and Rotnitzky and Powell (1990) [102]. Different from existing approaches, this test features stabilized variance estimation of facility effect estimates via an integrated correlation matrix shared by all patients with variable discharge counts. Despite being a score test, constrained model fitting under the null is unnecessary with the assumption that $\hat{\boldsymbol{\beta}}$

and $\hat{\boldsymbol{\eta}}$, estimated based on the entire sample of 541,769 discharges, are sufficiently accurate to replace $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$. Similar treatments can be found in the literature [49, 37, 36, 138].

The stabilized score test statistic T_i^{RS} under the null H_{0i} is given by

$$T_i^{\mathrm{RS}} := \frac{\mathcal{U}_i(\tilde{\boldsymbol{\gamma}}_i)}{\sqrt{\widetilde{V}_{\gamma_i}}}, \tag{2.7}$$

where $\mathcal{U}_i(\tilde{\boldsymbol{\gamma}}_i)$ is the ith element of the score vector $\mathcal{U}(\tilde{\boldsymbol{\gamma}}_i)$ with respect to facility effects $\boldsymbol{\gamma}$, $\widetilde{V}_{\gamma_i}$ is the middle piece of the stabilized robust variance estimator $\widetilde{\Sigma}_{\gamma_i}$ (defined in Appendix A.5) of the facility effect estimate, and tildes indicate evaluation at $\tilde{\boldsymbol{\theta}}_i := (\tilde{\boldsymbol{\gamma}}_i^\top, \hat{\boldsymbol{\eta}}^\top, \hat{\boldsymbol{\beta}})^\top$, with $\tilde{\boldsymbol{\gamma}}_i := (\hat{\gamma}_1, \ldots, \hat{\boldsymbol{\gamma}}_{i-1}, \hat{\gamma}_{\mathrm{M}}, \hat{\boldsymbol{\gamma}}_{i+1}, \ldots, \hat{\boldsymbol{\gamma}}_m)^\top$. The integrated correlation matrix shared by all patients is embedded in the variance estimator $\widetilde{V}_{\gamma_i}$. With the assumptions in Appendix A.4, as the number of patients in facility i approaches infinity, T_i^{RS} has an asymptotic standard normal distribution. Given a certain confidence level, a confidence interval for facility effect γ_i can be constructed by inverting (2.7).

2.4 Application

We apply our proposed methods to identify dialysis facilities with significantly better or worse performance than the national norm, leveraging the readmissions data derived from an extensive national ESRD patient database. Corresponding to a hospital discharge, each record includes patient demographics, clinical characteristics, and prevalent comorbidities for risk adjustment. Each discharge is followed by either a UHR or a competing event, including planned hospital readmission, death, and admission to a rehabilitation, psychiatric or long-term care hospital. Since facilities have little opportunity to oversee newly discharged patients, we exclude discharges with events over the first 3 days. Accordingly, the outcome of interest is defined as

an all-cause UHR to an acute care hospital within 4 to 30 days after discharge. In 2018, there were 272,897 patients in 6,937 Medicare-certified dialysis facilities with 541,769 qualifying discharges in total. These facilities had discharge counts ranging from 11 to 842 (mean 78.10), UHRs from 0 to 264 (mean 20.58), and competing events from 0 to 76 (mean 3.72). Further details regarding the data are available in Appendix A.8.

2.4.1 Implications of CRM on Profiling

With a logit link, we fit a CRM of 6,937 facility effects, 27 temporal (day) effects and 74 time-invariant covariates, and calculate the facility-specific SRRs, which range from 0 to 4.89, with quartiles 0.80, 1.00, and 1.20. For comparison, we also fit a logistic regression model (LRM) and present as scatter and box plots the log SRRs resulting from the two models in Figure 2.2, stratified by average time at risk, proportion of UHRs, and proportion of competing risks.

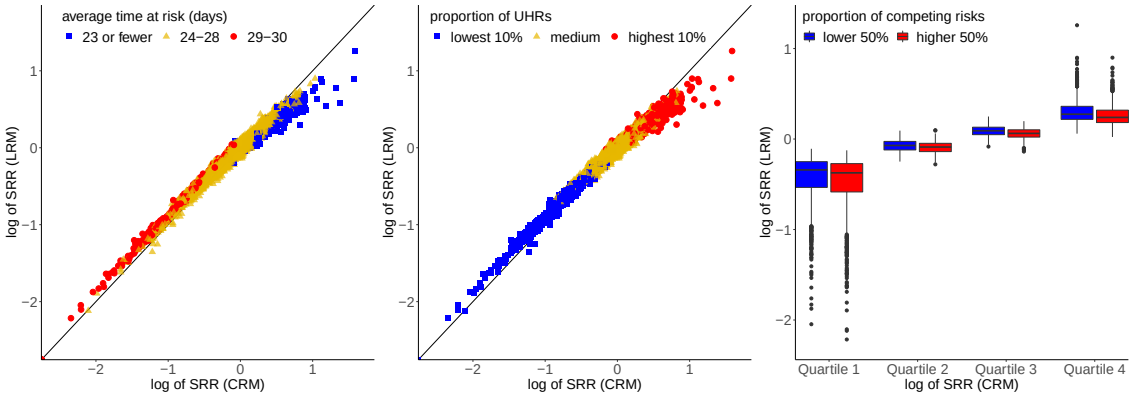


Figure 2.2: Scatter (left and middle panels) and box (right panel) plots of log SRRs under the LRM versus log SRRs under the CRM stratified by average time at risk (days), proportion of UHRs, and proportion of competing risks, respectively. SRRs under the CRM and LRM are computed based on (2.4) and He et al. [49], respectively. The at-risk time of a discharge is defined as the earlier of the time to the first event and the end of follow-up (30 days). 45-degree lines are in solid black in the left and middle panels. In the box plot (right panel), log SRRs from the CRM are grouped into quartiles.

Unsurprisingly, SRRs from the two models are positively correlated in the three panels. The CRM-based SRR tends to be greater than the LRM-based SRR

when the average time at risk is at most 23 days, and tends to be less than the LRM-based SRR when the at-risk time approaches the end of follow-up (Figure 2.2, left panel). This relationship holds as a consequence of the SRR definitions of the two models. At the discharge level, the contribution to the denominator of the LRM-based SRR equals ν in (2.5) evaluated at γ_{M}, [49] while according to (2.4), the contribution to the denominator of the CRM-based SRR equals the sum of the hazards of UHRs up to the at-risk time. This sum is easily seen to be greater than ν when the at-risk time approaches the end of follow-up, or be smaller than ν when the at-risk time is substantially shorter than the follow-up period. The relationship in the left panel then follows given that numerators of the two types of SRRs are the same. In light of this relationship, we conclude that although the LRM provides an appropriate description of the average performance of a facility over the course of follow-up, it leads to a shrinking the SRR toward the national average (i.e., SRR = 1), especially when a facility's average time at risk is relatively long or short. As standardized measures of the UHR burden within a facility, the LRM- and CRM-based SRRs both increase as the proportion of UHRs grows in the middle panel of Figure 2.2. As a side note, the first two panels of Figure 2.2 suggest that on average, a high-readmission facility tends to have shorter at-risk time than a low-readmission facility. The right panel of Figure 2.2 shows that within each quartile of the CRM-based SRR, the distribution of the LRM-based SRR shifts downward with a higher proportion of competing risks. This evidence confirms the inverse relationship between the LRM-based SRR and the rate of competing risks discussed in Section 2.2.2.

2.4.2 Score Tests with Different Variance Estimators

To examine the real-data behavior of the stabilized robust score test, we compare it with the classical robust [102] and model-based score tests (available in Appendix A.6). Given a certain significance level, a rejection of the null hypothesis H_{0i} with $\gamma_i > \gamma_{\mathrm{M}}$ (or $\gamma_i < \gamma_{\mathrm{M}}$) using one of the three tests indicates that the performance of facility i is significantly worse (or better) than the national average with regard to UHR, adjusting for risk factors. In Figure 2.3, we present the histograms and pairwise scatter plots of the test statistics with a significance level of 0.05. The diagonal panels with histograms reveal that the distribution of the classical robust score test statistics is more skewed than those of the other two, and it identifies 154 facilities as worse than expected and 519 as better, while the model-based score test flags 537 facilities as worse and 205 as better. The two non-stabilized tests both suffer from skewed outlier detection. In contrast, the stabilized robust test selects 352 worse and 218 better facilities, which is more conservative and balanced than the other two.

Pairwise scatter plots on off-diagonal panels break down one-way outlier counts into two-way tables. Specifically, 201 of the 6,264 non-outlying facilities under the classical robust test are flagged as worse by the stabilized robust test, while 309 normal facilities under the stabilized robust test are considered better than expected by the classical robust test. In addition, the dot distribution relative to the 45-degree line illustrates the differential levels of efficiency among the three score tests. In fact, 72.4% of facilities have their stabilized robust test statistic at least equal to their classical robust test statistic, 64.9% have their stabilized robust test statistic less than or equal to their model-based counterpart, and 76.5% have their classical robust test statistic no greater than their model-based test statistic. To summarize,

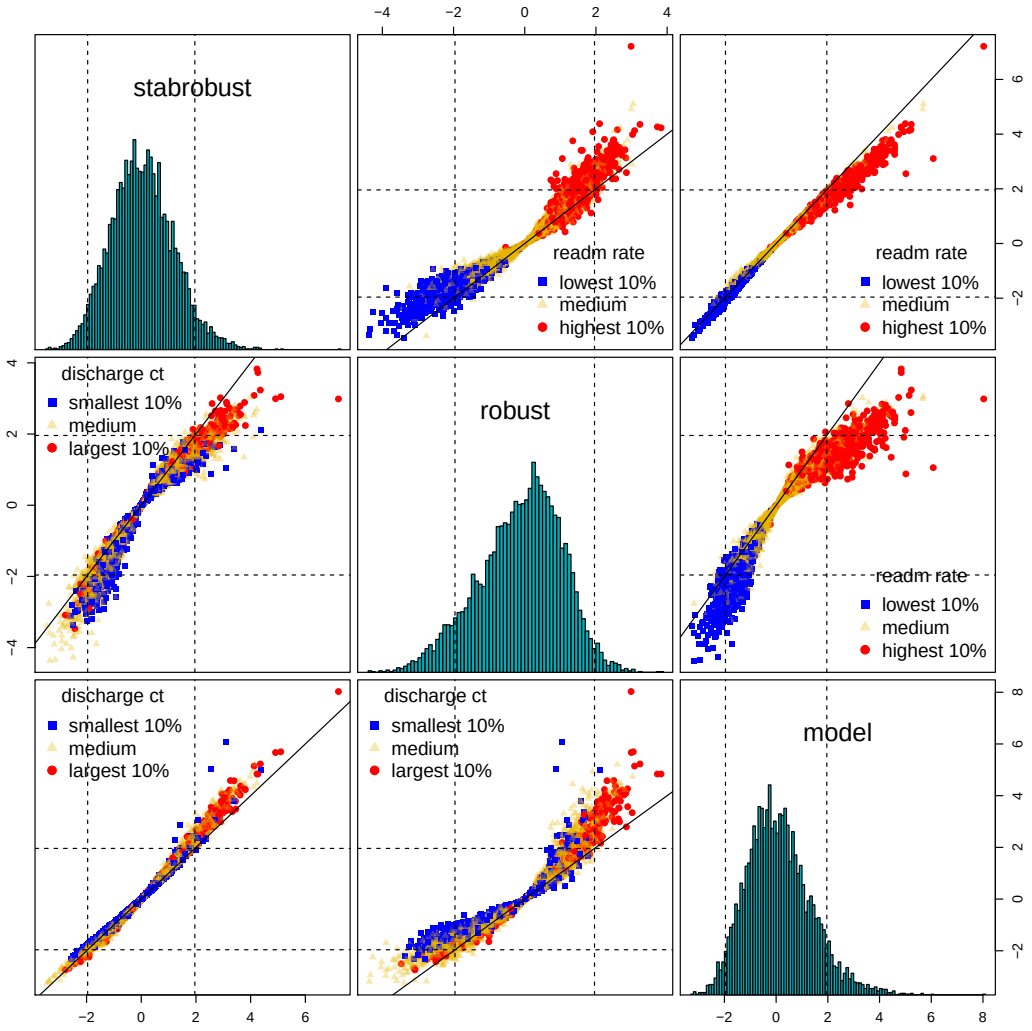


Figure 2.3: A matrix of histograms and scatter plots of score test statistics equipped with different variance estimators. "stabrobust", "robust" and "model" correspond to test statistics with the stabilized robust, classical robust and model-based variance estimators, respectively. Facilities are stratified by readmission rate or discharge count. Dashed lines represent 2.5% and 97.5% quantiles of the standard normal distribution. 45-degree lines are in solid black.

the stabilized robust score test has higher efficiency than the classical robust test. The efficiency advantage of the model-based score test is outweighed by its omission of the within-patient correlation.

2.4.3 Score versus Wald Tests

We compare score and Wald test statistics of facility effects in Figure 2.4, with three different variance estimators. Score tests outperform Wald tests with stable statistics and outlier detection. Specifically, Wald tests with the classical robust estimator (middle panel) have much smaller test statistics than those with the stabilized robust and model-based estimators (left and right). This evidence is consistent with the literature [40, 102, 8] in that the asymptotic distribution of a Wald test statistic is a poorer approximation to its small-sample distribution than that of a score test statistic, which can lead to inflated type I error.[124]

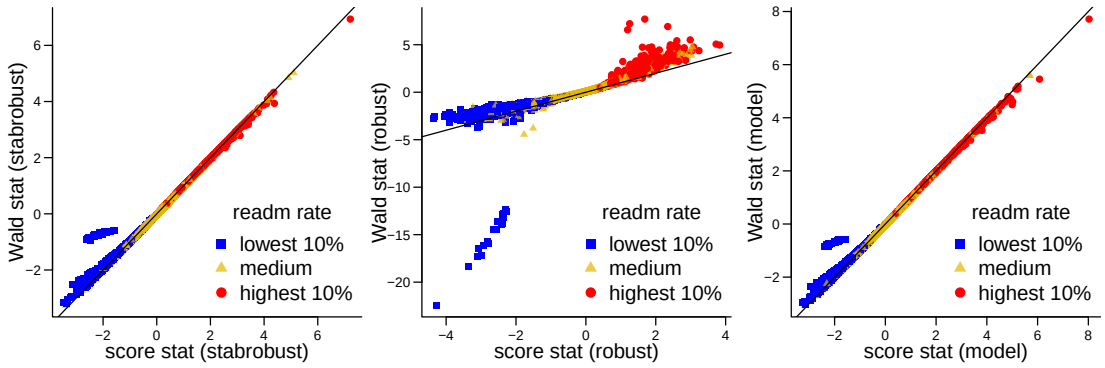


Figure 2.4: Scatter plots of Wald versus score test statistics with different variance estimators. Facilities are stratified by readmission rate. 45-degree lines are in solid black.

Besides superior small-sample performance, the stabilized robust score test (2.7) is readily available without refitting the CRM under the null. Note that the log-partial likelihood (2.3) is separable with respect to $\boldsymbol{\gamma}$, with the assumption that $\boldsymbol{\eta} \approx \hat{\boldsymbol{\eta}}$ and $\boldsymbol{\beta} \approx \hat{\boldsymbol{\beta}}$. Constrained maximization under $H_{0i} : \gamma_i = \gamma_{\mathrm{M}}$ thus does not change $\hat{\boldsymbol{\gamma}}$ except $\hat{\gamma}_i$, and (2.7) is easily computable by replacing $\hat{\gamma}_i$ with γ_{M}. As shown

in Figure A.4, using estimates from the constrained model refitting procedures, albeit theoretically valid, have hardly any discernible effect on the score test statistics, regardless of which variance estimator is considered.

2.5 Simulation Study

We present an application-driven simulation assessment of the stabilized robust score test, compared with the classical robust and model-based score tests. In each scenario, we create 1,000 data sets with $m = 1{,}000$ facilities. Facility-specific patient counts are drawn from Gamma(shape $= 2.589$, rate $= 0.066$), rounded up and left-truncated by 3. Patient-specific discharge counts are drawn from 1 to 12, with frequencies based on the real data. If the discharge count of a facility falls below 11, its first patient is assigned additional discharges. Each facility effect $\gamma_i \sim \mathcal{N}_1(0, \sigma_\gamma^2)$ with $\sigma_\gamma = 0.306$, and 27 temporal effects $\boldsymbol{\eta}$ are set based on the real data with mean -5.403 and standard deviation 0.155. Three discharge-specific covariates $\mathbf{Z}_{ipl} \sim \mathcal{N}_3(\mathbf{0}, \mathbf{I})$ with $\boldsymbol{\beta} = (1, 0.5, -1)^\top$. To introduce patient-level correlation, patient-specific random effects $\boldsymbol{\varepsilon}_{ip} \coloneqq (\varepsilon_{ip1}, \ldots, \varepsilon_{ipn_{ip}})^\top \sim \mathcal{N}_{n_{ip}}(\mathbf{0}, \boldsymbol{\Sigma})$ were added, where $\boldsymbol{\Sigma}$ is an exchangeable covariance matrix with marginal variance $\sigma^2 = 0.09$ and correlation ρ varying from 0 to 0.9. Further details are available in Appendix A.7.

UHRs and competing risks are sampled sequentially over time: Starting from Day 4, Bernoulli trials of UHRs with probabilities $h(\eta_k + \gamma_i + \mathbf{Z}_{ipl}^\top \boldsymbol{\beta} + \varepsilon_{ipl})$ under a logistic function h, and trials of competing risks with rate $5\%/27 = 0.185\%$ are applied to all discharges. Those with an event are marked as no longer at risk, and their event days are recorded as 4. From Day 5 until Day 30, trials are only applied to discharges at risk.

Table 2.1 displays type I error rates and powers, where we focus on Facility

1 by varying its discharge count n_1 and patient count $n^{(1)}$. Panel A shows error rates with 3 levels of facility size and 10 levels of ρ. When $\rho = 0$ (no correlation), the model-based score test has its error rates closest to the nominal level $\alpha = 0.05$, which suggests that the model-based test has higher efficiency than the other two conservative tests. When $\rho \neq 0$, the proposed test has error rates closest to 0.05 on average. The classical robust test is mostly conservative, especially for $n_1 = 11$, whereas the model-based test remains liberal. As expected, the error difference between the stabilized and classical robust test shrinks as facility size grows ($n_1 = 50$).

Panel B provides powers with 3 levels of facility size and 12 levels of relative deviation of facility effect γ_1/σ_γ, with $\rho = 0.5$. For all three tests, the power rises as facility size grows or absolute relative deviation enlarges, with faster power increase for positive deviation. The stabilized robust score test has greater power than the classical robust test, and greater power than the model-based test when $\gamma_1 < 0$ and $n_1 = 11$ or 20. To check whether the model-based test has inflated power, we reduce α from 0.05 to 0.04 for $n_1 = 50$ so that the type I error rates of the model-based test are close to those of the stabilized robust test with $\alpha = 0.05$. Given similar error rates, we observe that the model-based test has decreased powers with a mean of 0.675 across 12 levels of relative deviation. Since the average power of the stabilized robust test is 0.682 for $n_1 = 50$ and $\alpha = 0.05$, we conclude that the model-based score test suffers from inflated power.

2.6 Discussion

We propose a discrete competing risk model of readmission using a cause-specific hazards framework. Compared with existing logistic modeling approaches,

Table 2.1: Type I error rates and powers of score tests using different variance estimators. All values were calculated based on 1,000 independent replicates with $m = 1000$, $\sigma^2 = 0.09$, and significance level $\alpha = 0.05$. With correlation ρ varying from 0 to 0.9, rates in Panel A were obtained assuming $\gamma_1 = \gamma_M = 0$. In Panel B, correlation was fixed at $\rho = 0.5$, whereas γ_1 is allowed to vary in terms of relative deviation γ_1/σ_γ.

Panel A: Type I Error Rates									
ρ	$n_1 = 11$ and $n^{(1)} = 3$			$n_1 = 20$ and $n^{(1)} = 9$			$n_1 = 50$ and $n^{(1)} = 22$		
	stabrobust	robust	model	stabrobust	robust	model	stabrobust	robust	model
0	0.037	0.000	0.051	0.037	0.039	0.049	0.043	0.044	0.054
0.1	0.044	0.000	0.064	0.048	0.038	0.071	0.048	0.058	0.069
0.2	0.043	0.000	0.062	0.049	0.041	0.063	0.043	0.045	0.059
0.3	0.047	0.000	0.069	0.044	0.037	0.059	0.046	0.058	0.067
0.4	0.040	0.000	0.059	0.043	0.039	0.060	0.050	0.052	0.084
0.5	0.040	0.000	0.061	0.045	0.045	0.064	0.053	0.055	0.062
0.6	0.048	0.000	0.069	0.054	0.045	0.067	0.053	0.057	0.068
0.7	0.043	0.000	0.069	0.049	0.043	0.074	0.049	0.049	0.059
0.8	0.042	0.000	0.065	0.046	0.035	0.075	0.045	0.058	0.065
0.9	0.045	0.000	0.058	0.050	0.046	0.069	0.049	0.053	0.075
Panel B: Powers with $\rho = 0.5$									
$\frac{\gamma_1}{\sigma_\gamma}$	$n_1 = 11$ and $n^{(1)} = 3$			$n_1 = 20$ and $n^{(1)} = 9$			$n_1 = 50$ and $n^{(1)} = 22$		
	stabrobust	robust	model	stabrobust	robust	model	stabrobust	robust	model
- 4	0.180	0.000	0.123	0.340	0.231	0.316	0.787	0.721	0.801
-3.6	0.142	0.000	0.096	0.301	0.231	0.269	0.730	0.666	0.738
-3.2	0.135	0.000	0.095	0.279	0.190	0.241	0.623	0.569	0.634
-2.8	0.103	0.000	0.066	0.205	0.177	0.184	0.517	0.475	0.535
-2.4	0.078	0.000	0.048	0.162	0.128	0.148	0.443	0.417	0.458
-2	0.068	0.000	0.048	0.124	0.121	0.103	0.339	0.323	0.356
2	0.182	0.000	0.259	0.232	0.080	0.353	0.487	0.374	0.592
2.4	0.247	0.000	0.331	0.322	0.106	0.420	0.641	0.502	0.738
2.8	0.303	0.000	0.403	0.424	0.158	0.538	0.786	0.659	0.857
3.2	0.394	0.000	0.487	0.521	0.197	0.658	0.885	0.756	0.931
3.6	0.451	0.000	0.586	0.666	0.270	0.784	0.960	0.870	0.976
4	0.547	0.000	0.653	0.756	0.316	0.844	0.990	0.928	0.995

our model considers competing risks and event times, and leads to a more comprehensive approach to analyzing the outcome of interest. To facilitate estimation using the proposed model, we develop a fast-converging Blockwise Inversion Newton algorithm with scalability and memory efficiency compared to existing software packages. In addition, we devise a stabilized robust score test that improves the accuracy of inference in general, and is especially suitable for small facilities with low readmission rates. Evidence from simulations and application demonstrates the enhanced power, controlled type I error, and less skewed outlier detection of this test.

As in other survival contexts, incorporating event times into the analysis of readmissions data makes it possible to distinguish discharges of unequal lengths of risk exposure. In this regard, the traditionally used GLM framework is less comprehensive and may lead to spurious interpretation. For example, suppose that two facilities have the same rate of underlying (possibly unobserved) readmissions and that Facility 1 has more discharges followed by early deaths than Facility 2. Thus, Facility 2 would have a higher observed rate of readmission than Facility 1. Since the GLM ignores the times to death events, it would misinterpret Facility 1 as performing better than Facility 2, and lead to biased parameter estimation. Considering event times within a cause-specific hazard framework, Lee et al. [73] approached the parameter estimation by "naively" treating competing events as censorings. In contrast, our framework reparameterizes the discrete-time cause-specific hazard of competing risks, and decomposes the pseudo-likelihood function traditionally deemed as not factorizable. The proposed model is useful for analyzing a particular failure type of interest especially when the limited occurrence of nonignorable competing events hinders accurate joint modeling of all failure types.

The adoption of a discrete time framework has some advantages over the

more commonly encountered continuous time modeling in our application setting. The large-scale readmissions data consist of 541,769 discharges with only 27 distinct times. Thus, there are a large number of tied events at each time point. Using a discrete model in this context yields more accurate parameter estimation than using a continuous time model, which largely depends on approximation techniques to reduce estimation bias. In addition, the hazard formulation via a link function allows flexibility beyond the continuous Cox-type framework, in which the non-proportional hazards models can require painstaking effort on implementation and inference.

Growing out of readmission-focused provider profiling, our proposed framework is akin to mortality-based methods [63], and it can be used to evaluate providers according to alternative outcomes such as arteriovenous fistula access [109] and emergency department visits [61]. In addition, the competing risk model can be incorporated into cross-sectional and longitudinal provider monitoring techniques such as funnel plots [111] and cumulative sum control charts [112]. Although formulated with fixed facility effects, the underlying model can be readily implanted into a Bayesian framework with random facility effects. However, caution should be used in this case since the Bayesian or empirical Bayes approach can introduce substantial bias in regression parameters and effects for providers with extreme outcomes [49, 60, 57], which are those of primary interest.

Following the illness-death framework [139], Lee et al. [72, 71] and Haneuse and Lee [47] considered readmission as a nonterminal semi-competing risk and death as a terminal event. This semi-competing risk perspective facilitates explicit use of information on post-readmission death. Such models are useful and may well be worth further exploration, but are not applicable to the specific task of evaluating readmissions considered here. First, in our readmissions data, facility-specific death

rates were relatively low, making it impossible to employ a joint modeling framework to obtain stable and accurate parameter estimation for the death model. Second, by using a discharge- rather than a patient-level model, we also consider the possibility that a patient can have multiple readmissions, which was not explicitly considered in Lee et al [71].

CHAPTER III

Scalable Proximal Methods for Competing Risk Modeling with Time-Varying Coefficients

3.1 Introduction

The temporal variation in the effects of interventions or risk factors is a common phenomenon in time-to-event data [128, 120, 27]. To allow the effects to vary with time when analyzing the data, an important extension of the Cox model is often used—the relative risk model with time-varying coefficients. As remarked in Kalbfleisch and Prentice (2002) [59], this extended model is not only instrumental for testing the proportional hazards relationship, but also allows a concise description of a useful class of covariate effects. When the event of interest involves several distinct types, the time-varying effects can be similarly incorporated into a competing risk framework.

Our endeavors here were motivated by studying the cause-specific etiology of breast and prostate cancers using data from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) Program. Different from most analyses assuming constant effects of prognostic factors for survival, our purpose was to account for how the effects change with time. Early evidence from breast cancer patients [6, 5] suggested that tumor grade had a significant time-varying effect. As a more recent example, Brouwer (2020) [11] studied the cause-specific survival of pa-

tients diagnosed with squamous cell carcinomas (head and neck cancers) and found that the effects of age and sex were strongest at the time of diagnosis, but attenuated dramatically over the first few years. Ignoring the dynamic nature of a time-varying effect may weaken the internal validity of the study and cloud its implications for risk prediction, treatment development, and health care policy.

Along with the rising need for time-varying effect modeling in a cause-specific context, the growing volume and complexity of data pose overarching challenges to existing analytic frameworks. To name a few examples, Zucker and Karr (1990) [147] established a nonparametric penalized partial likelihood approach, which was revisited by Hastie and Tibshirani [48] with a cubic spline penalty. Gray [45, 46] instead used the cubic B-spline bases [24] with a small number of knots to parameterize the penalty function. Alternatively, Verweij and van Houwelingen (1995) [123] and Tutz and Binder (2004) [121] adopted as penalty the sum of squared pairwise differences of effect estimates at adjacent time points. In terms of implementation, these methods expand the original data in a repeated measurement format [116] using existing software such as the survival package [117], and perform well when the input data set is relatively small. As the data under analysis escalate in size, however, fitting a cause-specific hazard model with time-varying coefficients becomes formidably time-consuming and memory inefficient.

To illustrate this issue, we benchmarked the cause-specific hazard model fitting to simulated data sets (details in Section 3.5) using the function coxph of survival, called hereafter the Naive Newton (NaiveN) method. As shown in Figure 3.1, increasing the number of observations from 1,000 to 10,000 leads to substantial growth in the runtime and memory usage of NaiveN, whereas the runtime and memory consumption of our proposed algorithm slightly increase. If the sample size is further

scaled up to over 100,000, as in [11], even a well-configured workstation with 500 GB of RAM can barely accommodate the execution. The SEER breast cancer data we used consist of over 1 million patients, rendering any data-expansion-based method infeasible.

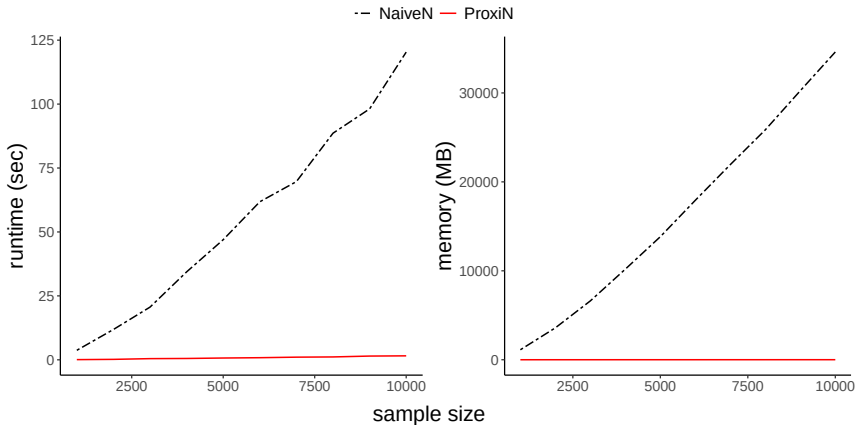


Figure 3.1: Runtime and memory usage of proximal Newton (ProxiN) and naive Newton (NaiveN) with sample sizes varying from 1,000 to 10,000. In each scenario, 10 data replicates were generated, and a fixed number of $K = 10$ knots were used for model fitting. Dichotomization was not applied to covariates. A tolerance level $\epsilon = 10^{-10}$ was used. The vertical axis displays average runtime (in seconds) across the 10 simulated data sets. Experiments were conducted on an Intel® Xeon® Gold 6254 quad-processor with max frequency 4 GHz and RAM 576 GB. ProxiN was implemented using Rcpp [33, 32] and RcppArmadillo [34]. Runtime and memory usage were measured using bench [52].

In the literature, some analyses have attempted to address this computational challenge. Inspired by a Kronecker product-based routine [93], He et al. (2017, 2021) [50, 51] respectively considered the Quasi-Newton (QuasiN) and minorize-maximization-based steepest ascent (MMSA) methods. Taking advantage of the large number of small strata in their settings, both methods demonstrated improved computation compared with the NaiveN, but were unable to handle an unstratified risk set with over one million subjects as in our cancer applications. Since gradient-based methods such as the MMSA only utilize first-order information, they often lead to appreciably more iterations than Newton-type methods. As will be seen in Table 3.1, the QuasiN may also produce highly biased estimates due to poor Hessian

matrix approximation.

In addition to the computational burden, numerical instability often arises from ill-conditioned second-order information in large-scale cause-specific hazard modeling. Specifically, when the data under analysis include a number of binary covariates with near-zero variation (e.g., in the SEER prostate cancer data, only 0.6% of the 716,553 patients had their tumors regional to the lymph nodes), the associated observed information matrix of a Newton-type method may have its minimum eigenvalue close to zero with a large condition number. Inverting such a nearly singular matrix is numerically unstable and the corresponding Newton updates are likely to be confined within a small neighborhood of the initial value, causing the estimates to be far from the optimal solutions. When multiple failure types are present, the issue of inaccurate estimation can be further exacerbated using existing methods (Section 3.5.1).

To achieve computational efficiency and reduce numerical instability, we propose a spline-based Newton-type method, which we term the proximal Newton (ProxiN) algorithm. This algorithm originates from the so-called proximal algorithms [91], and bears some resemblance to the more generic proximal Newton-type methods [69, 70]. Compared with the data-expansion-based NaiveN, the ProxiN reduces the execution time and memory consumption by orders of magnitude. As shown in Figure 3.1, the runtime and memory curves of the ProxiN stand in sharp contrast with those of the NaiveN and demonstrate the superiority of our proposed approach. Moreover, a shared-memory parallelization scheme further adds to the computational efficiency of the ProxiN with mild hardware requirements. As will be seen in Section 3.5.1, the ProxiN also leads to improved estimation accuracy compared to the NaiveN and QuasiN methods. The R and C++ code imple-

menting the ProxiN and the parallelization scheme is available online at https://github.com/UM-KevinHe/surtiver.

The rest of this chapter proceeds as follows: Section 3.2 lays out a cause-specific hazard model with time-varying coefficients. Section 3.3 presents the proximal Newton algorithm, its convergence properties, and the parallelization scheme. Section 3.4 introduces testing procedures. Simulation results are discussed in Section 3.5. In Section 3.6, the proposed method is applied to two large-scale cancer databases of SEER. Section 3.7 concludes with a discussion.

3.2 Model

For the ith subject, $i = 1, \ldots, n$, let T_i, C_i and $X_i \coloneqq T_i \wedge C_i$ denote the failure, censoring and observed time, respectively, where n denotes the total number of subjects and $a \wedge b \coloneqq \min\{a, b\}$. Let $\mathbf{Z}_i \coloneqq (Z_{i1}, \ldots, Z_{ip})^\top$ denote a vector of p covariates for risk adjustment. Let J_i be a random variable such that $J_i = j$ if subject i has a failure of type j, $j = 1, \ldots, m$, and $J_i = 0$ if subject i has a censoring event. Let $\Delta_{ij} \coloneqq I(T_i \leq C_i, J_i = j)$ be an indicator of type j failure, where $I(\cdot)$ is an indicator function. We assume that conditional on $\mathbf{Z}_i$, T_i is independently censored by C_i.

To model competing risks, we consider a Cox relative risk model

$$\lambda_j(t \mid \mathbf{Z}_i) \coloneqq \lambda_{0j}(t) \exp[\mathbf{Z}_i^\top \boldsymbol{\beta}_j(t)], \quad j = 1, \ldots, m, \tag{3.1}$$

where for failure type j, $\lambda_j(t \mid \mathbf{Z}_i)$ denotes the cause-specific hazard function, $\lambda_{0j}(t)$ denotes the baseline hazard, and $\boldsymbol{\beta}_j(t) \coloneqq [\beta_{j1}(t), \ldots, \beta_{jp}(t)]^\top$ is a p-dimensional vector of potentially time-varying coefficients. To estimate $\boldsymbol{\beta}_j(t)$ at time t, we span $\boldsymbol{\beta}_j(\cdot)$ by a set of K B-spline basis functions. Specifically, for $l = 1, \ldots, p$, $\beta_{jl}(\cdot)$ is

formulated as a linear combination

$$\beta_{jl}(t) \coloneqq \boldsymbol{\gamma}_{jl}^\top \mathbf{B}(t) = \sum_{k=1}^{K} B_k(t)\gamma_{jlk}, \tag{3.2}$$

where $\mathbf{B}(t) \coloneqq [B_1(t), \ldots, B_K(t)]^\top$ forms a basis, and $\boldsymbol{\gamma}_{jl} \coloneqq [\gamma_{jl1}, \ldots, \gamma_{jlK}]^\top$ is a vector of K unknown parameters for the lth time-varying coefficient $\beta_{jl}(\cdot)$ of failure type j. The time points at which pieces of B-spline polynomials meet are called knots and may be chosen based on the quantiles of the failure time points [45, 50, 51]. For ease of notation, we only consider a fixed number of K basis functions across different time-varying effects $\beta_{jl}(t)$; the general case of a varying number of basis functions is discussed in Section 3.5. Letting $\boldsymbol{\Gamma}_j \coloneqq [\boldsymbol{\gamma}_{j1}, \ldots, \boldsymbol{\gamma}_{jp}]^\top$, we define $\boldsymbol{\gamma}_j \coloneqq \text{vec}(\boldsymbol{\Gamma}_j^\top)$, a vectorization of $\boldsymbol{\Gamma}_j^\top$, by stacking its columns on top of each other, and $\boldsymbol{\gamma} \coloneqq [\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_m^\top]^\top$. Then model (3.1) leads to a log-partial likelihood given by

$$\ell(\boldsymbol{\gamma}) = \sum_{j=1}^{m} \ell_j(\boldsymbol{\gamma}_j), \tag{3.3}$$

in which

$$\begin{aligned} \ell_j(\boldsymbol{\gamma}_j) &\coloneqq \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij} \left[\mathbf{Z}_i^\top \boldsymbol{\Gamma}_j \mathbf{B}(X_i) - \log \left\{ \sum_{r \in R(X_i)} \exp\left(\mathbf{Z}_r^\top \boldsymbol{\Gamma}_j \mathbf{B}(X_i) \right) \right\} \right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij} \left[\mathbf{D}_i^\top(X_i)\boldsymbol{\gamma}_j - \log \left\{ \sum_{r \in R(X_i)} \exp\left(\mathbf{D}_r^\top(X_i)\boldsymbol{\gamma}_j \right) \right\} \right], \end{aligned} \tag{3.4}$$

where $R(X_i) \coloneqq \{r \in \{1, \ldots, n\} : X_r \geq X_i\}$ denotes the risk set of subject i, $\mathbf{D}_r(X_i) \coloneqq \mathbf{Z}_r \otimes \mathbf{B}(X_i)$, and $\otimes$ denotes the Kronecker product.

Observe that $\ell(\boldsymbol{\gamma})$ is twice continuously differentiable and concave since a log-sum-exp function is convex [9]. In addition, $\ell(\boldsymbol{\gamma})$ can be optimized by maximizing each $\ell_j(\boldsymbol{\gamma}_j)$ separately with respect to $\boldsymbol{\gamma}_j$. The gradient $\nabla \ell_j(\boldsymbol{\gamma}_j)$ and Hessian matrix $\nabla^2 \ell_j(\boldsymbol{\gamma}_j)$ of $\ell_j(\boldsymbol{\gamma}_j)$ are available in Appendix B.1.

3.3 Estimation

3.3.1 Proximal Newton algorithm

As discussed in Section 3.1, the classical Newton-type methods tend to provide unstable estimation, especially when the information matrix is nearly singular. Our proposed solution to this numerical instability has its roots in the proximal algorithm. For completeness, we start by reviewing this technique as well as its affinity to the traditional Newton approach. Interested readers are referred to Parikh and Boyd (2014) [91] for a detailed account.

Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be a closed and concave function; that is, its hypograph $\mathrm{hyp}(\ell) \coloneqq \{(\boldsymbol{\gamma}, s) \in \mathbb{R}^{d+1} : \ell(\boldsymbol{\gamma}) \geq s\}$ is a nonempty closed convex set. For any $\lambda > 0$, a proximal operator of $\lambda\ell$, denoted as $\mathbf{prox}_{\lambda\ell}$, is defined as

$$\mathbb{R}^d \ni \mathbf{v} \to \mathbf{prox}_{\lambda\ell}(\mathbf{v}) \coloneqq \operatorname*{argmax}_{\boldsymbol{\gamma}} \left\{ \ell(\boldsymbol{\gamma}) - \|\boldsymbol{\gamma} - \mathbf{v}\|_2^2/(2\lambda) \right\} \in \mathbb{R}^d, \tag{3.5}$$

where $\|\cdot\|_2$ denotes the Euclidean norm for vectors, or the induced L_2 norm for matrices. The use of argmax is justified by Proposition B.1.

To reveal the connection between the proximal operator (3.5) and Newton approach, note that if ℓ is twice continuously differentiable, its second-order Taylor approximation $\hat{\ell}_{\mathbf{v}}(\boldsymbol{\gamma})$ at $\mathbf{v}$ is $\hat{\ell}_{\mathbf{v}}(\boldsymbol{\gamma}) \coloneqq \ell(\mathbf{v}) + \nabla\ell^\top(\mathbf{v})(\boldsymbol{\gamma} - \mathbf{v}) + (\boldsymbol{\gamma} - \mathbf{v})^\top \nabla^2\ell(\mathbf{v})(\boldsymbol{\gamma} - \mathbf{v})/2$. To derive the proximal operator of $\lambda\widehat{\ell}_{\mathbf{v}}(\boldsymbol{\gamma})$, observe that the corresponding maximand is

$$\ell(\mathbf{v}) + \nabla\ell^\top(\mathbf{v})(\boldsymbol{\gamma} - \mathbf{v}) + (\boldsymbol{\gamma} - \mathbf{v})^\top \left(\nabla^2\ell(\mathbf{v}) - \mathbf{I}/\lambda \right) (\boldsymbol{\gamma} - \mathbf{v})/2,$$

where $\nabla^2\ell(\mathbf{v}) - \mathbf{I}/\lambda$ is a negative definite matrix with $\mathbf{I}$ being a $d \times d$ identity matrix. Maximizing the above quadratic maximand yields

$$\mathbf{prox}_{\lambda\widehat{\ell}_{\mathbf{v}}}(\mathbf{v}) = \mathbf{v} + \left(\mathbf{I}/\lambda - \nabla^2\ell(\mathbf{v}) \right)^{-1} \nabla\ell(\mathbf{v}), \tag{3.6}$$

which is a Levenberg–Marquardt step [74, 79], or a Newton step with a modified Hessian matrix [86].

As noted in Section 3.2, the log-partial likelihood $\ell(\boldsymbol{\gamma})$ in (3.3) is twice continuously differentiable and concave. Since a function is upper semi-continuous if and only if its hypograph is closed [101], (3.3) is also a closed function. Applying (3.6) to the second-order Taylor approximation of (3.3), we have the proximal Newton algorithm sketched as Algorithm 1, where $X_{j1} < \cdots < X_{jn_j}$ denote the n_j distinct times of type j failures, $j = 1, \ldots, m$, and $\mathbf{Z}_{jq}$ denotes the vector $\mathbf{Z}_i$ such that $\Delta_{ij} = 1$ and $X_i = X_{jq}$, $q = 1, \ldots, n_j$.

Algorithm 1: Proximal Newton Algorithm

for $j \leftarrow 1$ **to** m **do** ▷ m failure types
  initialize $s \leftarrow 0$, $\lambda_0 > 0$, and $\boldsymbol{\gamma}_j^{(0)} = \mathbf{0}$
  set $\phi \in (0, 0.5)$, $\psi \in (0.5, 1)$, $\delta \geq 1$ and $\epsilon > 0$
  do
    for $q \leftarrow 1$ **to** n_j **do** ▷ n_j distinct failure times
      for $u \leftarrow 0$ **to** 2 **do**
        $S_{jq}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq}) = \sum_{r \in R(X_{jq})} \exp\{[\mathbf{Z}_r \otimes \mathbf{B}(X_{jq})]^\top \boldsymbol{\gamma}_j^{(s)}\} \mathbf{Z}_r^{\odot u}$
      end for
      for $w \leftarrow 1$ **to** 2 **do**
        $\overline{\mathbf{Z}}_{jq}^{(w)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq}) = S_{jq}^{(w)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq}) / S_{jq}^{(0)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})$
      end for
      $\mathbf{V}_{jq}(\boldsymbol{\gamma}_j^{(s)}, X_{jq}) = \overline{\mathbf{Z}}_{jq}^{(2)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq}) - \left[\overline{\mathbf{Z}}_{jq}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})\right]^{\odot 2}$
    end for
    $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) = \frac{1}{n} \sum_{q=1}^{n_j} \left\{\mathbf{Z}_{jq} - \overline{\mathbf{Z}}_{jq}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})\right\} \otimes \mathbf{B}(X_{jq})$
    $\nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)}) = -\frac{1}{n} \sum_{q=1}^{n_j} \mathbf{V}_{jq}(\boldsymbol{\gamma}_j^{(s)}, X_{jq}) \otimes \{\mathbf{B}(X_{jq}) \mathbf{B}^\top(X_{jq})\}$
    $\Delta \boldsymbol{\gamma}_j^{(s)} = \left[\mathbf{I}/\lambda_s - \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)})\right]^{-1} \nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})$ ▷ Newton step
    $\eta^2 = \nabla \ell_j^\top(\boldsymbol{\gamma}_j^{(s)}) \Delta \boldsymbol{\gamma}_j^{(s)}$ ▷ η: Newton increment
    $\nu \leftarrow 1$
    while $\ell_j(\boldsymbol{\gamma}_j^{(s)} + \nu \Delta \boldsymbol{\gamma}_j^{(s)}) < \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \phi \nu \eta^2$ **do** ▷ line search
      $\nu \leftarrow \psi \nu$
    end while
    $\boldsymbol{\gamma}_j^{(s+1)} = \boldsymbol{\gamma}_j^{(s)} + \nu \Delta \boldsymbol{\gamma}_j^{(s)}$
    $\lambda_{s+1} = \delta \lambda_s$
    $s \leftarrow s + 1$
  while $\eta^2 \geq 2\epsilon$
end for

3.3.2 Convergence of the proximal Newton algorithm

The proposed proximal Newton algorithm, as a likelihood maximization approach, includes particular features to ensure convergence in most practical settings. First, the Newton step $\Delta\boldsymbol{\gamma}_j^{(s)}$ on Line 16 of Algorithm 1 is an ascent direction of $\ell_j(\boldsymbol{\gamma}_j^{(s)})$ at $\boldsymbol{\gamma}_j^{(s)}$, which is defined as follows:

Definition III.1. A direction $\boldsymbol{\mu} \in \mathbb{R}^d$ is an ascent direction of a function $\ell : \mathbb{R}^d \to \mathbb{R}$ at a point $\boldsymbol{\gamma} \in \mathbb{R}^d$ if $\exists\, \bar{\nu} > 0$ such that $\forall\, \nu \in (0, \bar{\nu}]$, $\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) > \ell(\boldsymbol{\gamma})$.

Using the concept of directional derivative, Definition III.1 implies that $\boldsymbol{\mu} \in \mathbb{R}^d$ is an ascent direction of a differentiable function ℓ at $\boldsymbol{\gamma}$ if

$$\lim_{\nu \to 0} \frac{\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) - \ell(\boldsymbol{\gamma})}{\nu} = \nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} > 0.$$

An equivalent condition is provided in the following Lemma III.2, which shows that $\Delta\boldsymbol{\gamma}_j^{(s)}$ on Line 16 is an ascent direction of $\ell_j(\boldsymbol{\gamma}_j^{(s)})$ at $\boldsymbol{\gamma}_j^{(s)}$ ($\mathbf{I}/\lambda_s - \nabla^2\ell_j(\boldsymbol{\gamma}_j^{(s)})$ is positive definite for any $\lambda_s > 0$). The proof of Lemma III.2 is available in Appendix B.2.

Lemma III.2. *Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then a direction $\boldsymbol{\mu} \in \mathbb{R}^d$ satisfies $\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} > 0$ at $\boldsymbol{\gamma}$ if and only if there exists a symmetric and positive definite matrix $\mathbf{M}$ such that $\boldsymbol{\mu} = \mathbf{M}^{-1}\nabla\ell(\boldsymbol{\gamma})$.*

Second, the backtracking line search on Lines 19 and 20 of Algorithm 1 constitutes a practical implementation of the Armijo–Goldstein conditions

$$\ell_j(\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}) \geq \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \phi\nu\nabla\ell_j^\top(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}, \tag{3.7}$$

$$\ell_j(\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}) \leq \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \psi\nu\nabla\ell_j^\top(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}, \tag{3.8}$$

$\phi \in (0, 0.5)$, $\psi \in (0.5, 1)$, based on which the step length ν is determined. Condition (3.7), known as the Armijo condition [2], explicitly requires a sufficient increase in

ℓ_j proportional to step length ν and directional derivative $\nabla\ell^\top(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}$ before the line search is terminated. However, (3.7) alone does not guarantee convergence since ϕ can be arbitrarily small. Condition (3.8), known as the Goldstein condition [42], imposes a lower bound on ν so that $\boldsymbol{\gamma}_j^{(s)}$ cannot be very close to $\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}$.

We present below three assumptions through which the convergence properties of the proximal Newton algorithm are achieved.

Assumption III.3. The log-partial likelihood component $\ell_j(\boldsymbol{\gamma}_j)$ of (3.4) is coercive, i.e., $\lim_{\|\boldsymbol{\gamma}_j\|_2\to\infty} \ell_j(\boldsymbol{\gamma}_j) = -\infty$, $j = 1, \ldots, m$.

As discussed in Lange (2013) [68], this assumption along with the continuity and concavity of ℓ_j guarantees that the superlevel set $\{\boldsymbol{\gamma}_j \in \mathbb{R}^{pK} : \ell_j(\boldsymbol{\gamma}_j) \geq \ell_j(\boldsymbol{\gamma}_j^{(0)})\}$ is convex and compact.

Assumption III.4. The matrix $\mathbf{I}/\lambda_s - \nabla^2\ell_j(\boldsymbol{\gamma}_j^{(s)})$ on Line 16 of Algorithm 1 has a bounded condition number, i.e., $\exists\,\kappa > 0$, such that

$$\mathbf{I}/\lambda_s - \nabla^2\ell_j(\boldsymbol{\gamma}_j^{(s)}) \leq \kappa, \quad j = 1, \ldots, m, \tag{3.9}$$

where for any invertible matrix $\mathbf{M}$, $\kappa_2(\mathbf{M}) \coloneqq \|\mathbf{M}\|_2\|\mathbf{M}^{-1}\|_2$.

Assumption III.5. The sequence $\{\lambda_s\}_{s=0}^{\infty}$ of positive tuning parameters monotonically approaches infinity as $s \to \infty$, i.e., $\lim_{s\to\infty} \lambda_s = \infty$.

The following theorem provides a set of convergence characterizations of Algorithm 1. The proof is included in Appendix B.2.

Theorem III.6. *Let ℓ_j assume* (3.4) *with an initial iterate $\boldsymbol{\gamma}_j^{(0)}$, and let $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ be a sequence of iterates defined by Line 22 of Algorithm 1, where $\Delta\boldsymbol{\gamma}_j^{(s)}$ is given by Line 16, and $\nu > 0$ is determined by* (3.7) *and* (3.8) *with $\phi \in (0, 0.5)$ and*

$\psi \in (0.5, 1)$. If Assumptions III.3 and III.4 hold, then $\{\ell_j(\boldsymbol{\gamma}_j^{(s)})\}_{s=0}^{\infty}$ converges and $\lim_{s\to\infty} \|\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})\|_2 = 0$.

Note that Theorem III.6 does not conclude with the convergence of $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=0}^{\infty}$. However, given the fact that $\boldsymbol{\gamma}_j^*$ is a global maximizer of the concave and differentiable function ℓ_j if and only if $\nabla \ell_j(\boldsymbol{\gamma}_j^*) = \mathbf{0}$, the ultimate iterate from Algorithm 1 should be close enough to the optimal solution with a sufficiently small tolerance ϵ in most practical situations.

With a priori assumptions on the optimal solution $\boldsymbol{\gamma}_j^*$, requiring $\phi \in (0, 0.5)$ and $\psi \in (0.5, 1)$ allows a step length ν equal to 1 to ultimately satisfy (3.7) and (3.8), and enables Algorithm 1 to achieve superlinear convergence as defined below. A formal statement is given in Theorem III.8, with the proof available in Appendix B.2.

Definition III.7. A sequence $\{\boldsymbol{\gamma}^{(s)}\}_{s=1}^{\infty} \subset \mathbb{R}^d$ converges superlinearly to $\boldsymbol{\gamma}^* \in \mathbb{R}^d$ if there exists a sequence $\{\xi_s\}_{s=1}^{\infty}$ of positive real numbers with $\lim_{s\to\infty} \xi_s = 0$ such that $\forall\, s \in \mathbb{N}$, $\|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2 \leq \xi_s \|\boldsymbol{\gamma}^{(s)} - \boldsymbol{\gamma}^*\|_2$.

Theorem III.8. *Let ℓ_j assume (3.4) with an initial iterate $\boldsymbol{\gamma}_j^{(0)}$, and let $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ be a sequence of iterates defined by Line 22 of Algorithm 1, where $\Delta\boldsymbol{\gamma}_j^{(s)}$ is given by Line 16, and $\nu > 0$ is determined by (3.7) and (3.8) with $\phi \in (0, 0.5)$ and $\psi \in (0.5, 1)$. In addition, assume that $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ converges to $\boldsymbol{\gamma}_j^*$ with a negative definite $\nabla^2 \ell_j(\boldsymbol{\gamma}_j^*)$. If Assumptions III.3 and III.5 hold, then (1) $\exists\, s_0 \in \mathbb{N}$ such that $\forall\, s \geq s_0$, $\nu = 1$ satisfies (3.7) and (3.8); (2) $\nabla \ell_j(\boldsymbol{\gamma}_j^*) = \mathbf{0}$; and (3) $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=0}^{\infty}$ converges superlinearly to $\boldsymbol{\gamma}_j^*$ provided that $\forall\, s \geq s_0$, $\nu = 1$ for some $s_0 \in \mathbb{N}$.*

3.3.3 Shared-memory parallelization

In the literature, various parallel computing schemes have been proposed to boost computational efficiency in generalized linear models (GLMs) [92, 30, 54], Bayesian inference [43], and random forests [133], among other instances. Despite the widespread recognition from the statistics community [31], there is a paucity of research on the application of parallel computing to large-scale time-to-event data, especially in a shared-memory context. The utmost reason is that modeling survival outcomes often involves risk-set-specific calculation tasks at all failure times. These tasks, unlike the observation-specific calculations in GLMs, are not equally costly in terms of computational complexity, since the size of the risk set $R(X_i)$ (defined in Section 3.2) varies with the observed time X_i. The unequal-sized risk sets resulting from an increasing sequence of failure times pose a challenge to load balancing, i.e., the distribution of tasks over a set of computing units (threads).

Following a distributed-memory framework, Lu et al. (2015) [78] bypassed this issue by sample stratification so that risk sets can only be formed within a certain stratum. However, their approach becomes infeasible if stratification is not possible. Moreover, as the sample size escalates (as in our cancer applications), the distributed-memory approach becomes less appealing since having multiple copies of a large data set concurrently is not memory-efficient.

In addition to load balancing arising from unequal-sized risk sets, the presence of time-varying coefficients poses a second challenge to parallel computing. When $\boldsymbol{\beta}_j(t)$ is time-invariant, i.e., $\boldsymbol{\beta}_j(t) = \boldsymbol{\beta}_j$, one may approach the problem by first calculating $\{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta}_j)\}_{i=1}^n$ and then obtaining the cumulative sums of $\{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta}_j)\}_{i=1}^n$ in parallel by means of the prefix sum algorithm [12]. When $\boldsymbol{\beta}_j(t)$ varies with time t, however, $\exp[\mathbf{Z}_i^\top \boldsymbol{\beta}_j(t)]$ has to be re-evaluated for different risk sets, making the

aforementioned approach infeasible.

To tackle the issue of load balancing in the presence of massive data and time-varying coefficients, we propose a shared-memory paradigm that optimizes workload allocation among a given number c of available threads where $c \geq 2$. For time X_{jq} of failure type j, let $n_{X_{jq}} \coloneqq |R(X_{jq})|$, i.e., the number of elements of the risk set $R(X_{jq})$. For failure type j, Algorithm 1 culminates in the calculations of $\ell_j(\boldsymbol{\gamma}_j^{(s)})$, $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})$ and $\nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)})$ at iteration s, which in turn depend upon $S_{jq}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})$. An analysis of time complexity reveals that computing $S_{jq}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, X_{jq})$ costs $O(pKn_{X_{jq}})$, $O(p(K+1)n_{X_{jq}})$ and $O(p(4K+3p+3)n_{X_{jq}})$, respectively, for $u = 0, 1, 2$. The linearity with respect to $n_{X_{jq}}$ suggests using as cutoffs the c-quantiles $\{\bar{n}_a\}_{a=1}^{c-1}$ of the cumulative sums of $\{n_{X_{jq}}\}_{q=0}^{n_j}$ (with $n_{X_{j0}} = 0$) to partition the collection of n_j risk sets into c subcollections of nearly equal computational costs.

Let $\bar{n}_c$ denote the sum of $\{n_{X_{jq}}\}_{q=0}^{n_j}$ and let $\bar{n}_0 = 0$. Algorithm 2 presents the parallelization of computing $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})$ at iteration s (Lines 5–14 of Algorithm 1), in which Line 8 is a race condition requiring execution on one thread at a time (nonparallel). The other two quantities can be obtained similarly. Evidence in Appendix B.4 using the SEER breast and prostate cancer data demonstrates the speedup and efficiency of the proposed parallelization scheme.

Algorithm 2: Parallel Computation of $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)})$ at Iteration s

1: initialize $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) \leftarrow \mathbf{0}$
2: **for** $a = 1$ **to** c **do in parallel** ▷ schedule c threads
3: **foreach** $b \in \{b : \bar{n}_{a-1} < \sum_{q=0}^{b} n_{X_{jq}} \leq \bar{n}_a\}$ **do** ▷ assign jobs to thread a
4: **for** $u \leftarrow 0$ **to** 1 **do**
5: $S_{jb}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, X_{jb}) = \sum_{r \in R(X_{jb})} \exp\{[\mathbf{Z}_r \otimes \mathbf{B}(X_{jb})]^\top \boldsymbol{\gamma}_j^{(s)}\} \mathbf{Z}_r^{\odot u}$
6: **end for**
7: $\overline{\mathbf{Z}}_{jb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, X_{jb}) = S_{jb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, X_{jb}) / S_{jb}^{(0)}(\boldsymbol{\gamma}_j^{(s)}, X_{jb})$
8: $\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) \leftarrow \nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \frac{1}{n} \left\{ \mathbf{Z}_{jb} - \overline{\mathbf{Z}}_{jb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, X_{jb}) \right\} \otimes \mathbf{B}(X_{jb})$ ▷ race
9: **end for**
10: **end for**

3.4 Hypothesis testing

Inferential attempts regarding the significance of the time-varying effects $\boldsymbol{\beta}_j(t)$ for type j failure can be formulated as the linear hypothesis $H_{01} : \mathbf{C}\boldsymbol{\beta}_j(t) = \mathbf{0}$, where $\mathbf{C}$ is a contrast matrix with full row rank r. Our penalty-free spline-based modeling and estimation lay the groundwork for a straightforward Wald-type significance test. By (3.2), the null H_{01} can be rewritten as $H_{01} : [\mathbf{C} \otimes \mathbf{B}^\top(t)]\boldsymbol{\gamma}_j = \mathbf{0}$, and a Wald test statistic is given by

$$\hat{\boldsymbol{\gamma}}_j^\top [\mathbf{C}^\top \otimes \mathbf{B}(t)] n \{[\mathbf{C} \otimes \mathbf{B}^\top(t)][\mathbf{I}/\lambda - \nabla^2 \ell_j(\hat{\boldsymbol{\gamma}}_j)]^{-1}[\mathbf{C}^\top \otimes \mathbf{B}(t)]\}^{-1}[\mathbf{C} \otimes \mathbf{B}^\top(t)]\hat{\boldsymbol{\gamma}}_j,$$

where $\hat{\boldsymbol{\gamma}}_j$ is the estimate of $\boldsymbol{\gamma}_j$. Under the null H_{01}, the test statistic approximately follows a chi-square distribution with r degrees of freedom. Pointwise confidence intervals across time are readily obtainable via test inversion. For instance, if one wants to test whether $\beta_{jl}(t) = 0$, where $\beta_{jl}(t)$ is the lth component of $\boldsymbol{\beta}_j(t)$, $l = 1, \ldots, p$, then $\mathbf{C} = [0, \ldots, 1, \ldots, 0]$, where only the lth element equals 1.

A second test of particular interest is to examine whether a certain effect $\beta_{jl}(t)$ is constant over time. In the literature, various procedures have been proposed to address this inference issue. As the default check for nonproportionality in the R package survival [117], Grambsch and Therneau (1994) [44] suggested a generalized least squares test on the scaled Schoenfeld residuals. Assuming $\beta_{jl}(t) = \beta_{jl} + \theta_{jl} g_{jl}(t)$ with unknown constants β_{jl} and θ_{jl}, and a possibly unknown function $g_{jl}(\cdot)$, the residuals are based on a one-step Newton-Raphson estimator $\hat{\theta}_{jl}$ of θ_{jl} and an estimator $\hat{\beta}_{jl}$ of β_{jl} from the Cox proportional hazards model. This approach provides a fast and easy check for nonproportionality without the need to fit a model of time-varying effects. Relying on a one-term Taylor approximation, however, using the scaled Schoenfeld residuals may lead to inflated type-I error when $|\beta_{jl}(t) - \beta_{jl}|$ is

large. In addition, the residual calculation may be unstable, particularly near the end of follow-up [118].

To test whether the effect $\beta_{jl}(t)$ is time-invariant, our approach amounts to a Wald test on the control points. Similar to He et al. (2017) [50], we observe that if $\gamma_{jl1} = \cdots = \gamma_{jlK} = \bar{\gamma}$, then

$$\beta_{jl}(t) = \bar{\gamma} \sum_{k=1}^{K} B_k(t) = \bar{\gamma},$$

in which we utilize the property of the B-spline basis that $\sum_{k=1}^{K} B_k(t) = 1$ for any t. This leads to the null hypothesis $H_{02l} : \bar{\mathbf{L}}\boldsymbol{\gamma}_{jl} = \mathbf{0}$, where $\bar{\mathbf{L}}$ is a $(K-1) \times K$ matrix given by

$$\bar{\mathbf{L}} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

A Wald test statistic can thus be constructed as

$$\hat{\boldsymbol{\gamma}}_{jl}^{\top} \bar{\mathbf{L}}^{\top} (\bar{\mathbf{L}} \mathbf{M}_{jl} \bar{\mathbf{L}}^{\top})^{-1} \bar{\mathbf{L}} \hat{\boldsymbol{\gamma}}_{jl},$$

where $\mathbf{M}_{jl}$ denotes the lth diagonal $K \times K$ block of $[\mathbf{I}/\lambda - \nabla^2 \ell_j(\hat{\boldsymbol{\gamma}}_j)]^{-1}/n$ for $l = 1, \ldots, p$. Under the null H_{02l}, this test statistic approximately follows a chi-square distribution with $K-1$ degrees of freedom.

Once the time-varying effects are distinguished from the time-independent ones through tests of nonproportionality, a cause-specific hazard model with time-variant and -invariant coefficients can be fit via an equality constrained maximization problem. Suppose $\beta_{jl_1}(t), \ldots, \beta_{jl_{\bar{p}}}(t)$ are flagged as time-variant effects. Let $\mathbf{L}$ be a $p(K-1) \times pK$ matrix whose lth $(K-1) \times K$ submatrix on the diagonal equals $\bar{\mathbf{L}}$ if $\beta_{jl}(t)$ is time-variant or $\mathbf{0}$ otherwise, and all off-diagonal blocks equal $\mathbf{0}$. Solving the

following problem

$$\begin{aligned} \underset{\Delta\boldsymbol{\gamma}_j \in \mathbb{R}^{pK}}{\text{maximize}} \quad & \nabla \ell_j^\top(\boldsymbol{\gamma}_j)\Delta\boldsymbol{\gamma}_j + \Delta\boldsymbol{\gamma}_j^\top[\nabla^2 \ell_j(\boldsymbol{\gamma}_j) - \mathbf{I}/\lambda]\Delta\boldsymbol{\gamma}_j/2 \\ \text{subject to} \quad & \mathbf{L}\Delta\boldsymbol{\gamma}_j = \mathbf{0}, \end{aligned}$$

in which $\boldsymbol{\gamma}_j$ is a feasible point satisfying $\mathbf{L}\boldsymbol{\gamma}_j = \mathbf{0}$ (e.g., $\boldsymbol{\gamma}_j = \mathbf{0}$), we can obtain the Newton step

$$\Delta\boldsymbol{\gamma}_j^* = \mathbf{U}\left[\mathbf{U}^\top\{\mathbf{I}/\lambda - \nabla^2 \ell_j(\boldsymbol{\gamma}_j)\}\mathbf{U}\right]^{-1}\mathbf{U}^\top \nabla \ell_j(\boldsymbol{\gamma}_j)$$

at each iteration (to replace Line 16 of Algorithm 1), where $\mathbf{U}$ is a $pK \times \bar{p}$ matrix, whose range (column space) is the null space of $\mathbf{L}$.

3.5 Simulation Study

To compare the proximal Newton algorithm with the NaiveN and QuasiN methods, we conducted a series of simulation experiments. The NaiveN was implemented via the function coxph in the R package survival, and the QuasiN was implemented using the base R function optim [86]. Since the estimation and inference with respect to different failure types can be handled separately within a cause-specific hazard framework, we focused primarily on a single failure type and dropped the index j to simplify notation.

In each simulation scenario, a number of independent data replicates were generated with the sample size n ranging from 1,000 to 10,000. We considered $p = 5$ covariates $\mathbf{Z}_i$ drawn from a multivariate normal distribution with zero mean, unit variance and an AR(1) correlation structure with parameter $\rho = 0.6$. To introduce numerical singularity, the continuous covariates were then dichotomized into binary variables, with the probability of being one uniformly varying from 0.8 to 0.9. This treatment intended to mimic our application setting where the Hessian matrix had a large condition number even when the number of observations was large. A constant

baseline hazard $\lambda_0(t) = 0.5$ was used with covariate parameters calibrated as $\boldsymbol{\beta}(t) = [1, \sin(3\pi t/4), -1, -1, 1]^\top$. Failure times were generated from the survivor function of (3.1), and censoring times were drawn from a uniform distribution between 0 and 3. Observed times were determined as the minimum of the failure and censoring time pairs.

3.5.1 Estimation accuracy

To assess the estimation accuracy of the proposed ProxiN, Table 3.1 presents the integrated mean squared error (IMSE), average bias, and average variance associated with the three algorithms. Model fitting was performed by treating all coefficients as time-dependent. Using a uniform distribution, we sampled 1,000 distinct time points from the interval between 0 and 3. At each time t, the mean estimates of $\beta_1(t)$ and $\beta_2(t)$ across 100 data replicates were used to calculate the mean squared error and variance, the difference of which is the squared bias. Taking the average across the 1,000 time points, we obtained the IMSE, average squared bias, and average variance. The average bias is simply the squared root of the average squared bias.

Panel A of Table 3.1 displays the three measures of estimation accuracy for $\beta_1(t)$. Since the nearly singular Hessian matrix was inaccurately approximated by a matrix in the BFGS algorithm, the QuasiN had consistently much higher IMSE than the other two methods. Of these two, the ProxiN had lower IMSE, bias and variance, especially when the sample size equaled 1,000 or 5,000. As a side observation, the IMSE was largely due to the variance component for all three methods. When it comes to the estimation accuracy of $\beta_2(t)$, the ProxiN overall outperformed the alternatives and the performance of QuasiN was even worse than that for $\beta_1(t)$. The difference in the accuracy measures among the first two approaches shrunk as

the sample size increased. To explore the impact of different censoring schemes, we varied the uniform distribution with different ranges of support (from $[0, 3]$ to $[1.5, 3]$), and used the exponential distribution with different rates (from 0.2 to 1.0) as an alternative scheme. In addition, we also considered the performance of the ProxiN in settings where the sample size was of a similar order as in our cancer applications. Results are also available in Appendix B.5.

As noted in Section 3.2, it is conceptually desirable that the number of B-spline basis functions is allowed to vary across different time-varying coefficients. Although a systematic investigation into such a general case is absent in the literature and beyond the scope of this article, we conducted simulation experiments (results available in Appendix B.5) to shed the first light on knot selection based on the variation of a covariate. The bottom line is that as the covariate variation shrinks toward zero, fewer knots should be applied to expanding a time-varying coefficient, so that the effect can be estimated with sufficient accuracy.

With the sample size equal to 1,000, Figure 3.2 displays the true value along with the pointwise mean of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ across 100 data replicates, where $\beta_4(t) = t^2 \exp(t/2)/9$ and $\beta_5(t) = \exp(-1.5t)$. The QuasiN was not included due to its poor performance. For $\beta_1(t)$, the estimate curve of ProxiN had much smaller deviance from the true value curve than the NaiveN, the deviance of which was in the opposite direction. As for the time-varying $\beta_2(t)$, the estimate curve of ProxiN varied closely along the true value curve, whereas the estimate curve of the NaiveN deviated from the true one when $t > 2$.

Given a 95% confidence level, Figure 3.3 compares the coverage probability (CP) of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ resulting from the ProxiN and NaiveN, with time t varying from 0 to 3. As time increases, the CP curve for $\hat{\beta}_1(t)$ from the ProxiN

Table 3.1: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated, and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$.

method	size	IMSE	bias	variance
		Panel A: $\beta_1(t)$		
ProxiN	1000	3.60	0.24	3.55
	5000	0.25	0.02	0.25
	10000	0.15	0.04	0.15
NaiveN	1000	35.82	0.99	34.84
	5000	0.26	0.03	0.26
	10000	0.15	0.05	0.15
QuasiN	1000	6772.12	69.37	1960.34
	5000	4870.17	40.94	3194.03
	10000	3969.22	44.47	1991.35
		Panel B: $\beta_2(t)$		
ProxiN	1000	1.82	0.28	1.74
	5000	0.18	0.20	0.14
	10000	0.14	0.23	0.09
NaiveN	1000	20.94	1.41	18.95
	5000	0.25	0.23	0.20
	10000	0.13	0.20	0.09
QuasiN	1000	72892.15	237.68	16400.41
	5000	41906.34	106.80	30499.53
	10000	26924.89	107.92	15279.30

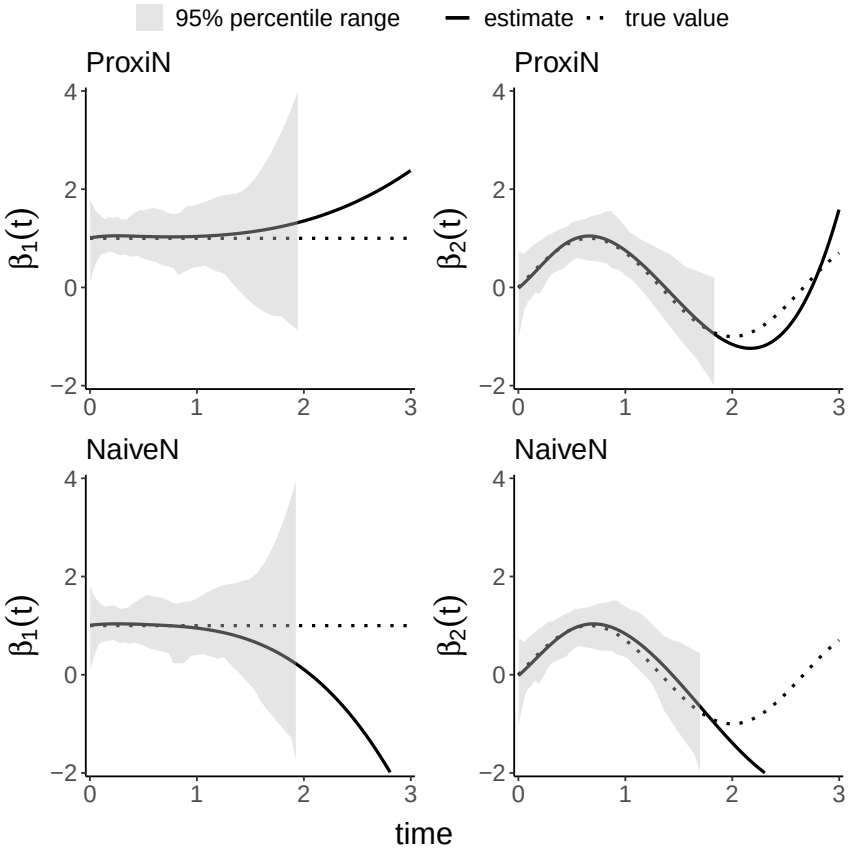


Figure 3.2: Mean of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ at each time t using the proximal Newton (ProxiN) and naive Newton (NaiveN) methods, with a 95% percentile range (2.5th and 97.5th percentiles as lower and upper limits). In each scenario, 100 data replicates were generated with sample size equal to 1,000. A fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$, with $\beta_3(t) = -1$, $\beta_4(t) = t^2 \exp(t/2)/9$, $\beta_5(t) = \exp(-1.5t)$.

algorithm fluctuates more closely around 0.95 than the NaiveN, though the CP curve of ProxiN drops sharply near the end of follow-up ($t = 3$) when $n = 5{,}000$ or 10,000. The QuasiN approach was not included as it often led to a singular Hessian matrix.

To illustrate the performance of ProxiN with more than one cause of failure, we compared the estimation accuracy of ProxiN, NaiveN and QuasiN with different sample sizes and two causes of failure (Table 3.2 and Table B.10). With the notation in Section 3.2, we set $\beta_{11}(t) = 1$, $\beta_{12}(t) = \sin(3\pi t/4)$, $\beta_{13}(t) = -1$, $\beta_{14}(t) = -1$, $\beta_{15}(t) = 1$ for the first failure type, and set $\beta_{21}(t) = -1$, $\beta_{22}(t) = \cos(3\pi t/4)$, $\beta_{23}(t) = 1$, $\beta_{24}(t) = 1$, $\beta_{25}(t) = -1$ for the second failure type. Failure times and types were determined based on Beyersmann et al. (2009) [7]. Censoring times were generated from a uniform distribution between 0 and 3. As in the case with only one cause of failure, the ProxiN outperformed the other two methods in terms of the IMSE, average bias, and average variance. A larger sample generally led to more accurate estimation of the true effects.

3.5.2 Testing for time-varying effects

The assessment of the test of nonproportionality is reported in Figure 3.4, where the average type-I error rate regarding a test of the time-invariant $\beta_1(t)$, and the average power regarding a test of the time-variant $\beta_2(t)$ across 1,000 data replicates are plotted against different levels of sample size, with a 5% significance level. When $\beta_2(t) = \sin(3\pi t/4)$ (top two panels), the ProxiN had a lower error curve for $\beta_1(t)$ and a higher power curve for $\beta_2(t)$. When the magnitude of $\beta_2(t)$ was tripled (bottom two panels), i.e., $\beta_2(t) = 3\sin(3\pi t/4)$, the NaiveN had much inflated error and power curves, both of which approached one as the sample size grew. By contrast, the proposed ProxiN maintained a controlled error curve around 5% as well as a high-level power line at one.

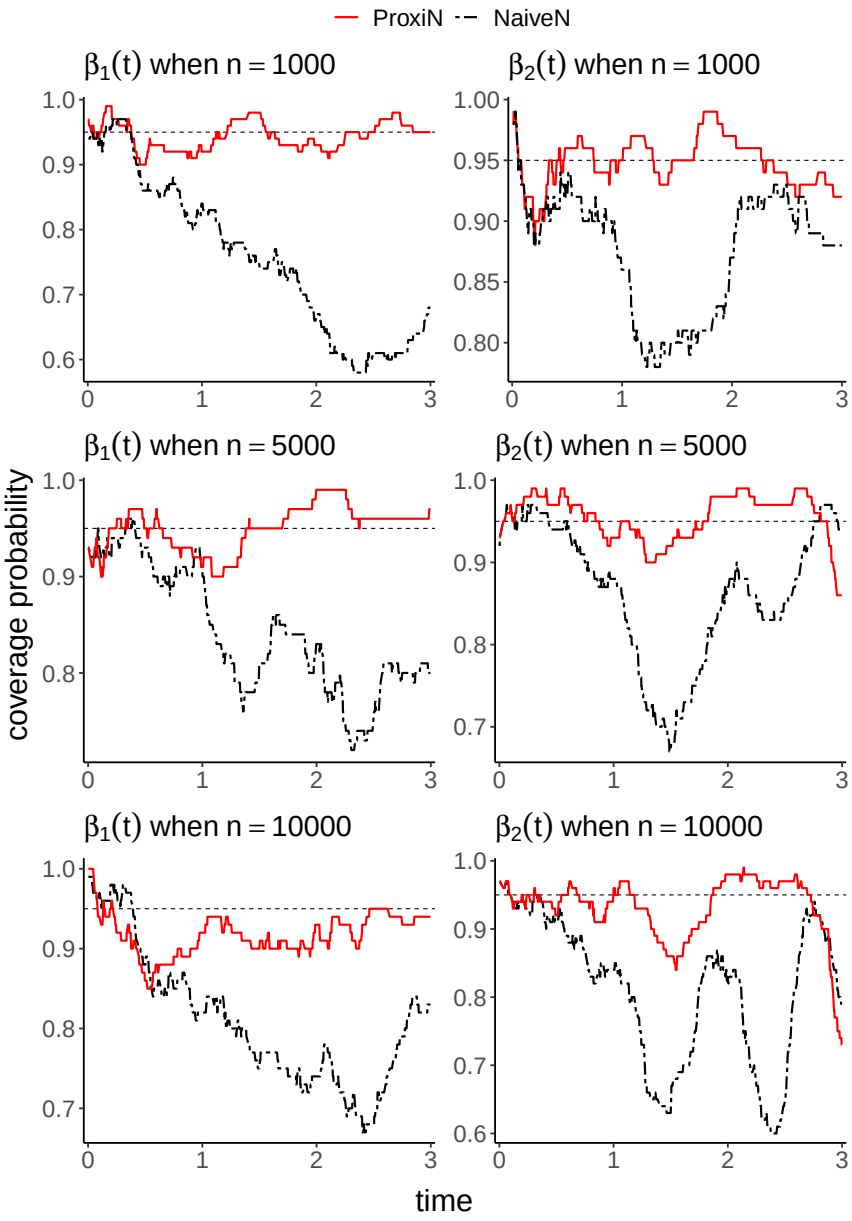


Figure 3.3: Coverage probability (CP) of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ at each time t using the proximal Newton (ProxiN) and naive Newton (NaiveN) methods, with a 95% confidence level. In each scenario, 100 data replicates were generated and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$.

Table 3.2: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_{11}(t)$ and $\hat{\beta}_{12}(t)$ (corresponding to the first cause of failure) using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_{11}(t) = 1$, $\beta_{12}(t) = \sin(3\pi t/4)$, $\beta_{13}(t) = -1$, $\beta_{14}(t) = -1$, and $\beta_{15}(t) = 1$.

method	size	IMSE	bias	variance
Panel A: $\beta_{11}(t)$				
ProxiN	1000	2.41	0.22	2.36
	5000	0.61	0.08	0.60
	10000	0.45	0.08	0.44
NaiveN	1000	7.72	0.68	7.26
	5000	3.75	0.06	3.74
	10000	2.63	0.35	2.51
QuasiN	1000	2830.04	41.82	1081.30
	5000	3715.09	34.98	2491.78
	10000	1700.60	28.43	892.08
Panel B: $\beta_{12}(t)$				
ProxiN	1000	2.47	0.22	2.42
	5000	1.02	0.25	0.96
	10000	0.71	0.17	0.68
NaiveN	1000	195.44	2.06	191.19
	5000	79.27	0.90	78.47
	10000	22.60	1.18	21.21
QuasiN	1000	111975.71	303.89	19627.88
	5000	61091.58	143.38	40532.74
	10000	17822.45	92.46	9274.30

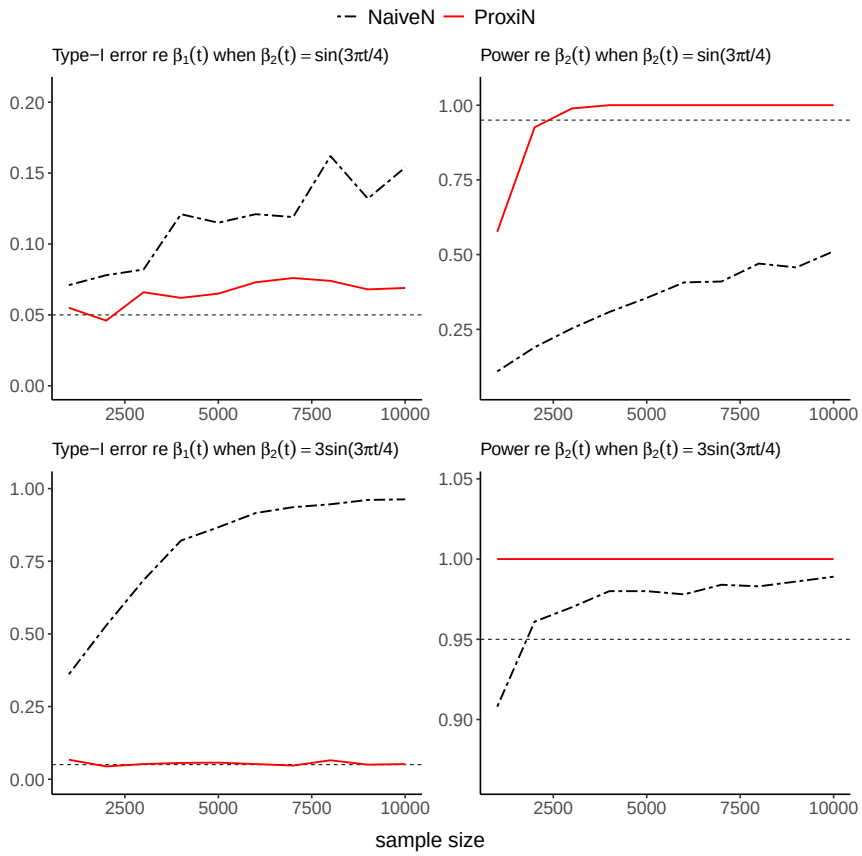


Figure 3.4: Type-I error rate and power regarding $\beta_1(t)$ and $\beta_2(t)$ using the proximal Newton (ProxiN) and naive Newton (NaiveN) methods with varying sample sizes. In each scenario, 1,000 data replicates were generated, and a fixed number of $K = 5$ knots were used for model fitting. In the first row, true values were $\beta_1(t) = 1$ and $\beta_2(t) = \sin(3\pi t/4)$, while in the second row, true values were $\beta_1(t) = 1$ and $\beta_2(t) = 3\sin(3\pi t/4)$.

3.6 Applications

To demonstrate the real-world performance of the proposed estimation and testing procedures, we applied these methods to the nationwide breast and prostate cancer survival database administered by the U.S. Surveillance, Epidemiology, and End Results (SEER) Program [114, 115].

3.6.1 SEER breast cancer data

For our study, 1,093,192 female patients first diagnosed with breast cancer between 1973 and 2015 were selected and their cause-specific deaths [11], if not censored, were recorded. In the analysis, we considered three risk factors: age, race and tumor stage at the time of diagnosis. Among all the patients, 24.21% were younger than 50 at diagnosis, 24.02% aged 50 to 59, 23.68% aged 60 to 69, and 28.09% were at least 70; 9.75% were black, 82.37% were white (including Hispanic), 7.42% belonged to other racial groups (American Indian, Alaska Native, Asian, Pacific Islander), and the remaining 0.46% were unknown. As for tumor staging, 60.02% had localized tumors, 31.39% had regionalized tumors, 6.13% had distant tumors, and 2.46% had their tumors recorded as unstaged. Event times (time to cancer death, other deaths or censoring) ranged from 1 month to 515 months, with a median of 80 months since diagnosis.

Treating cancer and other deaths as two distinct types of failure, we fit two cause-specific hazard models to the SEER breast cancer data with time-varying coefficients via Algorithm 1. Effect estimates as well as pointwise 95% confidence intervals are displayed in Figure 3.5 with a 20-year presentation. Treating the localized stage as the reference level and the other three as covariates, the top two panels display the overall shrinking staging effects on the two causes of death. Each of the three

stages had a larger effect on cancer death than that on other deaths. As expected, an advanced stage had a stronger effect on cancer death than an early stage. Relative to the white cohort, black breast cancer patients were more likely to die as a result of either cancer or other causes. They had an initial increase in the hazard of cancer death, followed by a gradual decrease to nearly zero. In contrast, the shrinkage of race effects on other deaths was slower. The three effects of age groups on cancer death immediately declined after diagnosis and then either remained stable (older than 70) or gradually increased (younger than 60). Age effects on other deaths remained relatively flat as time passed. The speedup and efficiency of the parallelized ProxiN is discussed in detail in Appendix B.3.

3.6.2 SEER prostate cancer data

In the prostate cancer data, 716,553 patients with a first diagnosis of prostate cancer between 2004 and 2017 were chosen and their cause-specific deaths or censorings were recorded. Similarly as in the analysis of breast cancer data, we examined age, race and tumor stage at the time of diagnosis. Among all the patients, 2.79% were younger than 50 at diagnosis, 20.83% aged 50 to 59, 40.74% aged 60 to 69, and 35.64% were at least 70; 14.58% were black, 69.44% were non-Hispanic white, 8.81% were Hispanic, and the remaining 7.17% belonged to other racial groups. (Since this data were collected only starting in 2004, the registry used different ethnic groupings than the breast cancer data, which started in 1973.) In terms of summary staging, 82.41% had localized tumors, 11.32% had regionalized tumors by direct extension, 0.6% had regional tumors to lymph nodes, 1.12% had their tumors as regional both by direct extension and lymph nodes, and 4.54% had tumors of unknown stage. Event times ranged from 1 month to 167 months, with a median of 6 years since diagnosis.

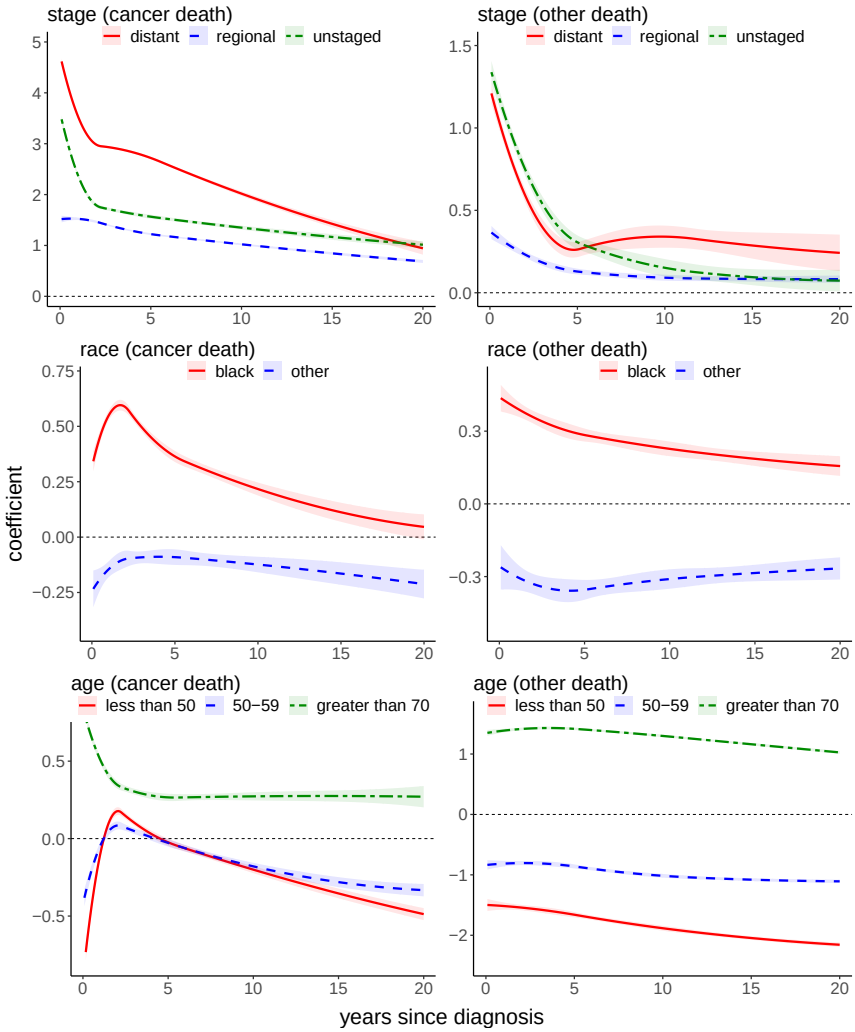


Figure 3.5: Estimates of the time-varying effects of tumor stage, race and age on death (due to cancer or other causes) as a function of time since diagnosis using the SEER breast cancer data. Quadratic B-splines were applied throughout the analysis with $K = 5$ knots. The ribbons in all panels represent 95% pointwise confidence intervals for the time-varying coefficients. At a 5% level, all effects on cancer death or other deaths were significantly time-dependent using the testing procedure in Section 3.4.

As in the application of breast cancer, we fit two cause-specific hazard models with time-varying coefficients to the SEER prostate cancer data. Estimates and confidence intervals are displayed in Figure 3.6 with a 10-year presentation. With the localized stage as the reference group and the other four as covariates, the top two panels reveal different patterns of staging effects on the two types of death. Overall, an advanced tumor stage led to a considerably higher hazard ratio of cancer death than the hazard ratio of other deaths. While the effects of regional both and regional by direct extension on cancer death were significantly positive, their effects on other deaths were negative. Nonproportionality tests with 5% size of the staging effects on cancer death indicated that they should all be viewed as time-variant. Relative to the white cohort, black prostate cancer patients were more likely to die as a result of either cancer or other causes. As expected, older patients had a higher hazard of dying from any cause than younger patients.

3.7 Discussion

The increasing availability of large-scale and complex data has the potential to vastly improve our understanding of important real-world problems such as cancer survival, but only with methodological and computational advances. Existing data-expansion- or gradient-based methods impose formidable computational costs and numerical instability to model fitting. To facilitate efficient and accurate statistical analysis in this context, we propose the proximal Newton algorithm along with a shared-memory parallelization paradigm and testing procedures. Simulation analyses demonstrate superior scalability, efficiency and estimation accuracy compared to alternative approaches. Applications to the SEER breast and prostate cancer data confirm the excellent real-world performance of our proposed approach.

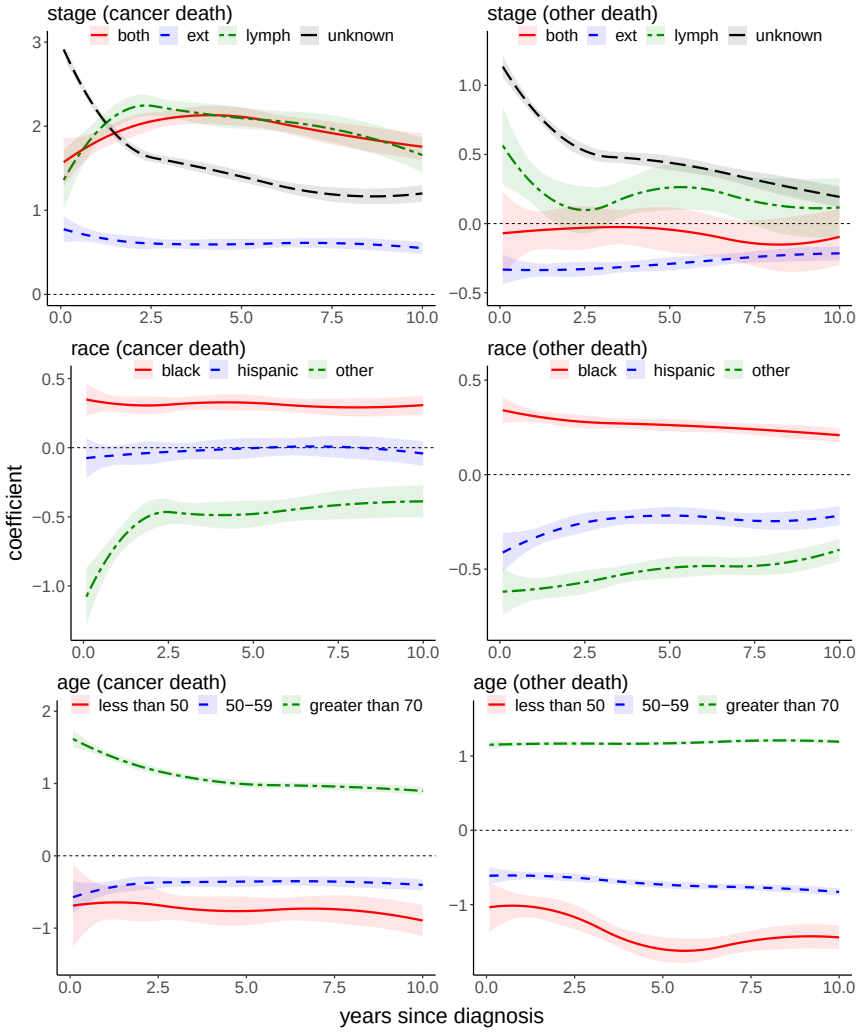


Figure 3.6: Estimates of the time-varying effects of tumor stage, race and age on death (due to cancer or other causes) as a function of time since diagnosis using the SEER prostate cancer data. Quadratic B-splines were applied throughout the analysis with $K = 5$ knots. The ribbons in all panels represent 95% pointwise confidence intervals for the time-varying coefficients. The four stages displayed in the legends are regional both by direct extension and lymph nodes (both), regional by direct extension (ext), regional by lymph nodes (lymph) and unknown. At a 5% level, significant time-varying effects on cancer death included age greater than 70, other races and the four stage effects. All effects on other deaths were significantly time-dependent except both and lymph.

Although developed for analyzing cancer data, the proposed technique can be used in many other applications that involve time-varying effect analysis. In kidney transplantation, for example, the relative risk of death among recipients relative to those on dialysis is known to initially increase due to surgery, but the subsequent decrease eventually leads to an overall survival benefit [128]. Similarly, when comparing two infant feeding strategies for preventing mother-to-child human immunodeficiency virus transmission, evidence from a randomized trial showed that, although breastfeeding with prophylaxis was associated with lower infant mortality at 7 months relative to formula feeding, this difference shrunk to insignificance through age 18 months [120]. Obesity, a well-known risk factor of mortality in the general population, was found among dialysis patients to have a short-term protective effect on survival and an increased risk of death after a long-term exposure [56, 55, 26, 27]. In all these instances, our proposed methods would have undoubtedly contributed to a better understanding of the changes in effects over time.

Depending on specific analytic needs, the proximal Newton algorithm can also be applied to a more general setting with stratum-specific baseline hazards. In a head and neck cancer application, for instance, there was evidence of substantial differences in the baseline hazards by tumor stage [11]. A stratified analysis taking account of the stage-wise variation may better reflect the effect evolution of prognostic factors. As another example, the analysis of electronic health records often involves integrating data from multiple health care providers. Stratification by providers can alleviate the mediation between provider-specific effects and the effects of risk factors. In either case, our proposed method can readily handle the less demanding computational burdens with reduced risk sets.

As for the determination of the number and location of knots in the cause-

specific hazard model, we followed the rules by Gray (1992) [45], that is, a small number of knots (e.g., 10) chosen to include an equal number of events within each time interval. Although using this early suggestion yields stable estimation in our applications, a systematic guideline on this issue is beyond the current endeavors. In addition, it is worth further exploration into the use of the penalized B-spline to alleviate overfitting and increase smoothness in coefficient estimation. Moreover, when the dimension of the parameter space is very high, existing model selection techniques such as Yan and Huang (2012) [140] would no longer be feasible. This necessitates in-depth investigation into high-dimensional variable selection methods with time-varying effects. Fortunately, the superb performance of the proposed algorithm paves the way for possible advances along these paths in a large-scale cause-specific setting.

In the top right panel of Figure 3.6, the effect curve of lymph on other deaths has more variation than the curve of ext especially for the initial 2.5 years since diagnosis, but the test of nonproportionality identified the effect of ext as time-dependent rather than the effect of lymph. This suggests that the effect of lymph on other deaths may not be nonzero everywhere. Although addressing this issue systematically is beyond the aims of the current article, more analytical effort is worthwhile on accounting for zero-effect regions in competing risk models with time-varying effects. Currently, there is a paucity of studies in the survival literature on time-varying effect modeling with zero-effect regions. For a relevant account on varying coefficients with zero-effect regions in the context of generalized linear models, we refer to a recent work by Yang (2020) [141].

CHAPTER IV

Understanding the Dynamic Impact of COVID-19 through Competing Risk Modeling with Bivariate Varying Coefficients

4.1 Introduction

This chapter grows out of our investigation in response to the request by the U.S. Centers for Medicare & Medicaid Services (CMS) on the influence of the coronavirus disease 2019 (COVID-19) pandemic on patients with end-stage renal disease (ESRD) [135]. Our goal is to inform evidence-based COVID-19 adjustment in the implementation of ESRD quality measures, especially for postdischarge patient outcomes. These quality measures have been routinely reported on Care Compare–Dialysis Facilities [18] to assess dialysis facilities in the ESRD Quality Incentive Program [19]. The calculation of pandemic-adjusted ESRD quality metrics largely depends on how COVID-19 as a risk factor should be accounted for in statistical modeling; any switch in measure-based flagging (e.g., from average to worse than expected) resulting from COVID-19 adjustment would lead to a substantial change in performance-based payments to dialysis facilities. This significant consequence indicates the high-stakes nature of our statistical endeavors.

To understand the impact of COVID-19 on patients requiring routine kidney dialysis for appropriate risk adjustment in CMS reporting, we explored their post-

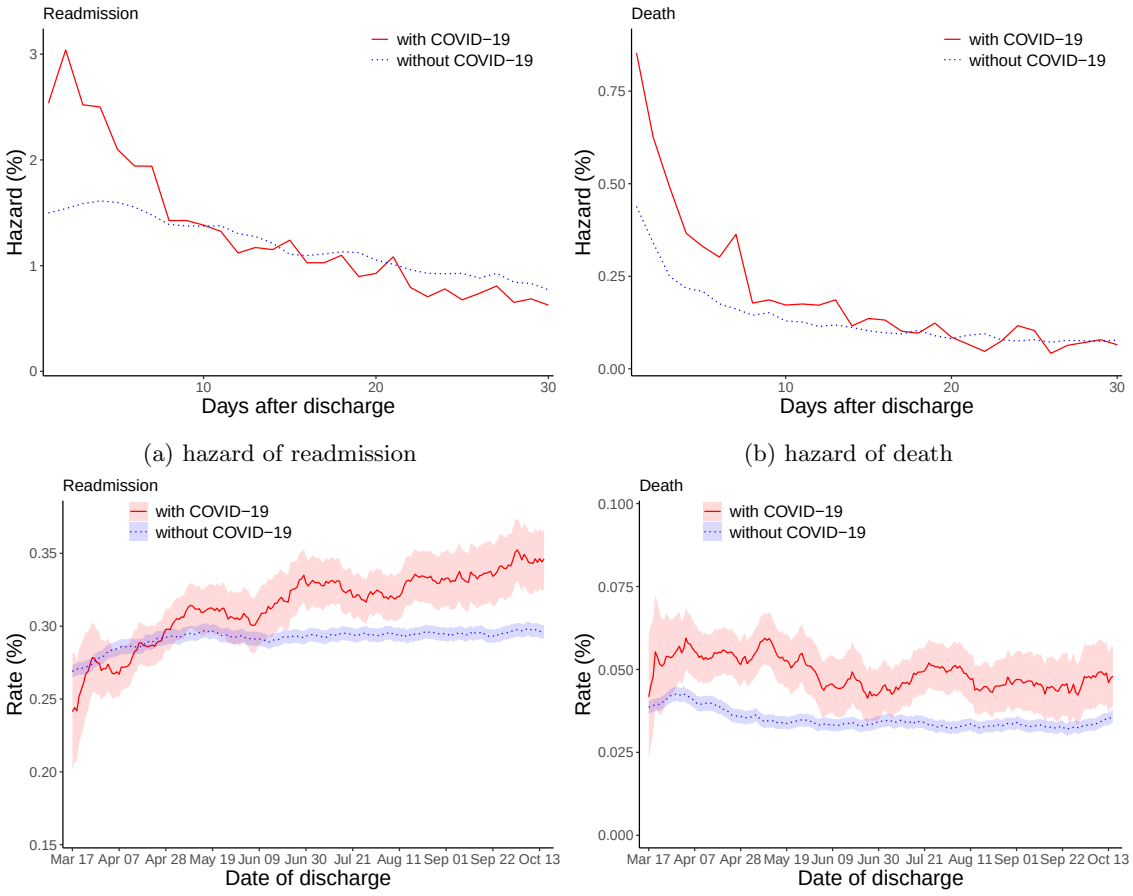
discharge readmissions and deaths by in-hospital COVID-19 diagnosis (with versus without COVID-19). Included in the data were 436,745 live hospital discharges of 222,154 Medicare dialysis beneficiaries from 7,871 dialysis facilities throughout the first ten months of 2020. The top two panels of Figure 4.1 shows that within a week of hospital discharge, the descending (unadjusted) cause-specific hazard curves of readmission and death were substantially higher for the group with COVID-19 than the group without. Figure 4.1c shows that the rate of readmission increased in both groups between mid-March and mid-May; from early June onward, the rate of readmission among discharges with COVID-19 began to significantly surpass the rate of readmission among discharges without. Figure 4.1d indicates that the rate of death in both groups started at a relatively high level and then overall decreased until mid-October; the rate of death among discharges with COVID-19 remained significantly higher than the rate among discharges without throughout the ten months. Despite the fact that other risk factors were not adjusted for, these preliminary findings indicate that the impact of COVID-19 was constantly changing with both postdischarge (Figures 4.1a and 4.1b) and calendar time (Figures 4.1c and 4.1d).

Existing risk adjustment models for quality measure development and health care provider monitoring mostly treat the outcome of interest as a binary variable, using logistic regression with fixed and random effects to indicate inter-provider variation [49, 60, 37, 36, 137, 87, 88, 3, 80]. Because these models do not account for event timing, they cannot be applied to our COVID-19 study to discover the important evidence of postdischarge variation; a time-to-event modeling framework would better meet the analytical needs in this setting. In addition, the unusual dynamics of the COVID-19 effect calls for a distinctive varying coefficient model that provides a unified characterization of the significant variations with both postdischarge and

calendar time, and a systematic inferential procedure testing the two-dimensional variations either jointly or separately. Unfortunately, such a flexible and comprehensive model is still lacking in the statistical literature.

Motivated by the pressing need for novel statistical methods to appropriately analyze the dynamic impact of COVID-19, we develop a spline-based bivariate varying coefficient model, treating postdischarge readmission and death as competing risks within a cause-specific hazard framework. Unlike existing time-varying coefficient models for time-to-event outcomes [147, 45, 48, 123, 121, 50, 51, 136], the proposed model formulates the effect of a risk factor (e.g., in-hospital COVID-19 diagnosis in our applications) as a bivariate function of both event time (e.g., postdischarge time to a readmission or death, hereafter postdischarge time) and an external covariate (e.g., calendar time since pandemic onset, hereafter calendar time). Tensor-product B-splines [106] are employed to estimate the surface of the bivariate COVID-19 effect, thereby allowing complex variation trajectories along two different dimensions. Although tensor-product B-splines were previously used to model interactions between two continuous risk factors [45], our study is the first to use this technique to characterize the complexly varying effect of a risk factor in a competing risk analysis.

Fitting the bivariate varying coefficient model to the massive postdischarge outcome data for Medicare dialysis patients poses significant computational issues that no existing methods can handle. Current methods rely on expanding a single observation into multiple records from the baseline until the observed time [116]. With large-scale data, this approach leads to prolonged convergence and overloaded memory even for univariate time-varying effect modeling [136], let alone bivariate varying coefficients. Moreover, the presence of extremely distributed binary covariates often



(a) hazard of readmission

(b) hazard of death

(c) rate of readmission

(d) rate of death

Figure 4.1: Panels (a) and (b) present unadjusted cause-specific hazard curves of unplanned hospital readmission and death, respectively, from January 1, 2020 to October 31, 2020. On each postdischarge day, the unadjusted hazard of readmission or death was defined as the number of readmissions or deaths occurring over that day divided by the number of discharges at risk for readmission (or death) at the beginning of that day. Panels (c) and (d) present rates of unplanned hospital readmission and death, respectively, among discharges with and without in-hospital COVID-19 from March 17, 2020 to October 15, 2020. Monthly rates and their 95% confidence intervals were calculated on a rolling basis.

introduce numerical instability with ill-conditioned Hessian matrices. To address these challenges, we develop a tensor-product proximal Newton algorithm that optimizes the unpenalized log-partial likelihood. This algorithm efficiently extends the approach by Wu et al. (2022) [136] to a two-dimensional setting. Leveraging the property of B-splines, we propose a hypothesis testing framework with respect to both univariate and bivariate variation of the COVID-19 effect.

To mitigate model overfitting and the wiggliness of the estimated COVID-19 effect surface in a multivariate setting, we also introduce difference-based anisotropic penalization [130, 131, 35] to the original log-partial likelihood, where the penalization is applied against the deviation from a constant coefficient model, and the degree of the penalty is regulated through dimension-specific sets of tuning parameters. The asymptotic distribution of the resulting penalized estimates is investigated under mild conditions, and a corresponding inference procedure that generalizes the test of Gray (1992) [45] is developed. To determine optimal tuning parameters, we evaluate various methods of cross-validation and extend the method of cross-validated deviance residuals to the setting with varying coefficients.

The rest of this chapter is organized as follows: Section 4.2 introduces the bivariate varying coefficient model for competing risks. Section 4.3 presents estimation and inference methods based on the unpenalized partial likelihood. In Section 4.4, we develop estimation and inference methods based on the penalized partial likelihood. Next, we demonstrate and evaluate the proposed methods with simulation experiments in Section 4.5 and two applications to Medicare dialysis patients in Section 4.6. Section 4.7 concludes with a discussion.

4.2 Model

First, we present a competing risk model with bivariate varying coefficients. For the ith subject in the gth stratum ($g = 1, \ldots, G$, $i = 1, \ldots, n_g$, where n_g denotes the total number of subjects in the gth stratum, i.e., dialysis facility in our applications), let T_{gi}, C_{gi}, and $X_{gi} \coloneqq \min\{T_{gi}, C_{gi}\}$ denote the failure, censoring and observed times, respectively. Let $\mathbf{Z}_{gi}$ denote a vector of p covariates associated with p bivariate varying coefficients, and let $\mathbf{W}_{gi}$ denote a vector of q covariates with invariant coefficients. For ease of notation and due to the interest of our applications, we assume that all bivariate varying coefficients depend upon a single effect modifying covariate $\breve{X}_{gi}$, although the dependence can be easily relaxed to be coefficient-specific. Further, let J_{gi} be a random variable such that $J_{gi} = j$ ($j = 1, \ldots, m$) if subject i in stratum g has a failure of type j, and $J_{gi} = 0$ if that subject is censored. In our applications (details in Section 4.6), j indicates different postdischarge outcomes (unplanned hospital readmission and death) or discharge destinations (to home, to another health care facility, and in-hospital death or to hospice). Let $\Delta_{jgi} \coloneqq I(T_{gi} \leq C_{gi}, J_{gi} = j)$ indicate whether subject i in stratum g has a type j failure, where $I(\cdot)$ is an indicator function, and let $\Delta_{gi} \coloneqq I(T_{gi} \leq C_{gi})$. We assume that conditional on $\mathbf{Z}_{gi}$, $\mathbf{W}_{gi}$ and $\breve{X}_{gi}$, T_{gi} and C_{gi} are independent so that the censoring is non-informative.

We consider a stratified Cox relative risk model with semi-varying coefficients [38], i.e.,

$$\lambda_{jgi}(t \mid \mathbf{Z}_{gi}, \mathbf{W}_{gi}, \breve{X}_{gi}) \coloneqq \lambda_{0jg}(t) \exp\left\{\mathbf{Z}_{gi}^{\top}\boldsymbol{\beta}_j(t, \breve{X}_{gi}) + \mathbf{W}_{gi}^{\top}\boldsymbol{\theta}_j\right\}, \tag{4.1}$$

where $\lambda_{jgi}(t \mid \mathbf{Z}_{gi}, \mathbf{W}_{gi}, \breve{X}_{gi})$ denotes the stratum- and cause-specific hazard function for failure type j, $\lambda_{0jg}(t)$ denotes the baseline hazard function allowed to be arbitrary

and assumed completely unrelated, $\boldsymbol{\beta}_j(t, \breve{X}_{gi}) \coloneqq [\beta_{j1}(t, \breve{X}_{gi}), \ldots, \beta_{jp}(t, \breve{X}_{gi})]^\top$ is a p-dimensional vector of varying coefficients, each of which is a bivariate function of time t and covariate $\breve{X}_{gi}$, and $\boldsymbol{\theta}_j$ is a q-dimensional vector of invariant coefficients. In our setting, t denotes the time (in days) since hospital discharge or admission, and $\breve{X}_{gi}$ denotes the discharge or admission time (in days) since the onset of the COVID-19 pandemic.

To approximate the surface of $\beta_{jl}(t, \breve{x})$, $l = 1, \ldots, p$, we span $\beta_{jl}(\cdot, \cdot)$ by tensor-product B-splines. Specifically,

$$\beta_{jl}(t, \breve{x}) \coloneqq \breve{\mathbf{B}}^\top(\breve{x})\boldsymbol{\gamma}_{jl}\mathbf{B}(t) = \sum_{\breve{k}=1}^{\breve{K}} \sum_{k=1}^{K} \gamma_{jl\breve{k}k} \breve{B}_{\breve{k}}(\breve{x}) B_k(t), \tag{4.2}$$

where $\mathbf{B}(t) \coloneqq [B_1(t), \ldots, B_K(t)]^\top$ and $\breve{\mathbf{B}}(\breve{x}) \coloneqq [\breve{B}_1(\breve{x}), \ldots, \breve{B}_{\breve{K}}(\breve{x})]^\top$ are B-spline bases (with intercept terms) at t and $\breve{x}$, respectively, and $\boldsymbol{\gamma}_{jl} \coloneqq [\gamma_{jl\breve{k}k}]$ is a $\breve{K} \times K$ matrix of unknown control points for the lth bivariate varying coefficient $\beta_{jl}(\cdot, \cdot)$ of failure type j. The number K (or $\breve{K}$) of B-spline functions forming a basis $\mathbf{B}(t)$ (or $\breve{\mathbf{B}}(\breve{x})$) relates to the degree d (or $\breve{d}$) of the piecewise B-spline polynomials and to the number u (or $\breve{u}$) of interior knots in that $K = u + d + 1$ (or $\breve{K} = \breve{u} + \breve{d} + 1$) [106]. In reality, interior knots of the B-spline space $\mathbf{B}(\cdot)$ can be chosen based on the quantiles of distinct failure times $\{X_{gi} : \Delta_{gi} = 1, i = 1, \ldots, n_g, g = 1, \ldots, G\}$ [45, 50, 51, 136], and the interior knots of $\breve{\mathbf{B}}(\cdot)$ can be set at the quantiles of covariates $\{\breve{X}_{gi} : i = 1, \ldots, n_g, g = 1, \ldots, G\}$.

Note that (4.2) can be rewritten as

$$\beta_{jl}(t, \breve{x}) = \{\mathrm{vec}(\boldsymbol{\gamma}_{jl}^\top)\}^\top \{\breve{\mathbf{B}}(\breve{x}) \otimes \mathbf{B}(t)\},$$

where vec denotes the vectorization of a matrix, i.e., stacking columns of a matrix on top of one another, and $\otimes$ denotes the Kronecker product. It follows that

$$\boldsymbol{\beta}_j(t, \breve{x}) = \boldsymbol{\Gamma}_j \{\breve{\mathbf{B}}(\breve{x}) \otimes \mathbf{B}(t)\},$$

where $\boldsymbol{\Gamma}_j \coloneqq [\text{vec}(\boldsymbol{\gamma}_{j1}^\top), \ldots, \text{vec}(\boldsymbol{\gamma}_{jp}^\top)]^\top$. Let $\boldsymbol{\gamma}_j \coloneqq \text{vec}(\boldsymbol{\Gamma}_j^\top)$, $\boldsymbol{\gamma} \coloneqq [\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_m^\top]^\top$, and $\boldsymbol{\theta} \coloneqq [\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_m^\top]^\top$. Given model (3.1), we have the log-partial likelihood

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \ell_j(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j) = \sum_{j=1}^{m} \sum_{g=1}^{G} \ell_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j), \tag{4.3}$$

in which

$$\ell_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j) = \sum_{i=1}^{n_g} \Delta_{jgi} \left[\mathbf{L}_{gi}^\top(X_{gi})\boldsymbol{\gamma}_j + \mathbf{W}_{gi}^\top \boldsymbol{\theta}_j - \log \left\{ \sum_{r \in R_g(X_{gi})} \exp\left(\mathbf{L}_{gr}^\top(X_{gi})\boldsymbol{\gamma}_j + \mathbf{W}_{gr}^\top \boldsymbol{\theta}_j \right) \right\} \right], \tag{4.4}$$

$R_g(X_{gi}) \coloneqq \{r \in \{1, \ldots, n_g\} : X_{gr} \geq X_{gi}\}$ denotes the risk set of subject i in stratum g, and $\mathbf{L}_{gr}(X_i) \coloneqq \mathbf{Z}_{gr} \otimes \breve{\mathbf{B}}(\breve{X}_{gr}) \otimes \mathbf{B}(X_{gi})$. The gradient and Hessian matrix of $\ell_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ are available in Appendix C.1.

4.3 Unpenalized partial likelihood approach

4.3.1 Estimation

As noted before, the joint estimation of the bivariate varying coefficient functions $\beta_{jl}(\cdot, \cdot)$ and invariant coefficients $\boldsymbol{\theta}_j$ based on the unpenalized log-partial likelihood (3.3) becomes computationally challenging, especially when the sample includes at least half a million subjects. To address this challenge, we develop a tensor product proximal Newton algorithm on the basis of Wu et al. (2022) [136] to allow bivariate varying coefficient estimation. This approach is derived from the proximal operator [91] of the second-order Taylor approximation of the log-partial likelihood (3.3), leading to a modified Hessian matrix. The algorithm features accurate and efficient model fitting to large-scale competing risks data with millions of subjects and binary predictors of near-zero variance. Let $X_{jg1} < \cdots < X_{jgn_{jg}}$ denote the n_{jg} distinct times of type j failures within stratum g. For failure time X_{jgb}, $b = 1, \ldots, n_{jg}$, let $\mathbf{Z}_{jgb}$, $\mathbf{W}_{jgb}$, and $\breve{X}_{jgb}$ denote $\mathbf{Z}_{gi}$, $\mathbf{W}_{gi}$, and $\breve{X}_{gi}$, respectively, such that $\Delta_{jgi} = 1$ and

$X_{gi} = X_{jgb}$. The algorithm is outlined as Algorithm 3. For theoretical arguments justifying the convergence of the algorithm, the reader is referred to Wu et al. (2022) [136]. In what follows, we will use carets to indicate unpenalized estimates resulting from this algorithm. For instance, $\hat{\boldsymbol{\gamma}}_{jl}$ denotes unpenalized estimates of $\boldsymbol{\gamma}_{jl}$.

Algorithm 3: Tensor Product Proximal Newton

for $j \leftarrow 1$ **to** m **do** ▷ m failure types
  initialize $s \leftarrow 0$, $\lambda_0 > 0$, $\boldsymbol{\gamma}_j^{(0)} = \mathbf{0}$, and $\boldsymbol{\theta}_j^{(0)} = \mathbf{0}$
  set $\phi \in (0, 0.5)$, $\psi \in (0.5, 1)$, $\delta \geq 1$ and $\epsilon > 0$
  do
    for $g \leftarrow 1$ **to** G **do** ▷ G distinct strata
      for $b \leftarrow 1$ **to** n_{jg} **do** ▷ n_{jg} distinct failure times
        for $u \leftarrow 0$ **to** 2 **do**
          $S_{jgb}^{(u)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) = \sum_{r \in R_g(X_{jgb})} \exp\{\mathbf{L}_{gr}^\top(X_{jgb})\boldsymbol{\gamma}_j^{(s)} + \mathbf{W}_{gr}^\top \boldsymbol{\theta}_j^{(s)}\} \begin{bmatrix} \mathbf{L}_{gr}(X_{jgb}) \\ \mathbf{W}_{gr} \end{bmatrix}^{\odot u}$
        end for
        for $w \leftarrow 1$ **to** 2 **do**
          $\mathbf{U}_{jgb}^{(w)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) = S_{jgb}^{(w)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) / S_{jgb}^{(0)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb})$
        end for
        $\mathbf{V}_{jgb}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) = \mathbf{U}_{jgb}^{(2)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) - \left[\mathbf{U}_{jgb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb})\right]^{\odot 2}$
      end for
    end for
    $\dot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) = \sum_{g=1}^{G} \sum_{q=1}^{n_j} \left\{ \begin{bmatrix} \mathbf{L}_{jgb}(X_{jgb}) \\ \mathbf{W}_{jgb} \end{bmatrix} - \mathbf{U}_{jgb}^{(1)}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb}) \right\}$
    $\ddot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) = -\sum_{g=1}^{G} \sum_{q=1}^{n_j} \mathbf{V}_{jgb}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}, X_{jgb})$
    $\begin{bmatrix} \Delta\boldsymbol{\gamma}_j^{(s)} \\ \Delta\boldsymbol{\theta}_j^{(s)} \end{bmatrix} = \left[\mathbf{I}/\lambda_s - \ddot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)})/n\right]^{-1} \dot{\ell}_j(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)})/n$ ▷ Newton step
    $\eta^2 = \dot{\ell}_j^\top(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) \begin{bmatrix} \Delta\boldsymbol{\gamma}_j^{(s)} \\ \Delta\boldsymbol{\theta}_j^{(s)} \end{bmatrix}$ ▷ η: Newton increment
    $\nu \leftarrow 1$
    while $\sum_{g=1}^{G} \ell_{jg}(\boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)} + \nu\Delta\boldsymbol{\theta}_j^{(s)}) < \sum_{g=1}^{G} \ell_{jg}(\boldsymbol{\gamma}_j^{(s)}, \boldsymbol{\theta}_j^{(s)}) + \phi\nu\eta^2$ **do** ▷ line search
      $\nu \leftarrow \psi\nu$
    end while
    $\boldsymbol{\gamma}_j^{(s+1)} = \boldsymbol{\gamma}_j^{(s)} + \nu\Delta\boldsymbol{\gamma}_j^{(s)}$
    $\boldsymbol{\theta}_j^{(s+1)} = \boldsymbol{\theta}_j^{(s)} + \nu\Delta\boldsymbol{\theta}_j^{(s)}$
    $\lambda_{s+1} = \delta\lambda_s$
    $s \leftarrow s + 1$
  while $\eta^2 \geq 2\epsilon$
end for

4.3.2 Inference

To examine the dynamic impact of COVID-19 among dialysis patients, it is logical to test whether a bivariate coefficient $\beta_{jl}(t, \breve{x})$ varies significantly with t and $\breve{x}$, either separately or jointly. By the property of B-splines, when $\gamma_{jl\breve{k}k}$ remains constant with k, i.e., $\gamma_{jl\breve{k}k} \equiv \gamma_{jl\breve{k}\cdot}$ for any $\breve{k} = 1, \ldots, \breve{K}$, $\beta_{jl}(t, \breve{x})$ reduces to

$$\beta_{jl}(t, \breve{x}) = \sum_{\breve{k}=1}^{\breve{K}} \gamma_{jl\breve{k}\cdot} \breve{B}_{\breve{k}}(\breve{x}) \sum_{k=1}^{K} B_k(t) = \sum_{\breve{k}=1}^{\breve{K}} \gamma_{jl\breve{k}\cdot} \breve{B}_{\breve{k}}(\breve{x}),$$

which no longer varies with t due to the fact that $\sum_{k=1}^{K} B_k(t) = 1$. This relationship suggests the null hypothesis $H_0^{(t)} : \mathbf{C}^{(t)} \mathrm{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$ for testing whether $\beta_{jl}(t, \breve{x})$ varies significantly with t, where $\mathbf{C}^{(t)} = \mathrm{diag}(\underbrace{\mathbf{D}, \ldots, \mathbf{D}}_{\breve{K}})$ is a block diagonal matrix with $\breve{K}$ diagonal blocks. Each of these blocks is a $(K-1) \times K$ first-order difference matrix $\mathbf{D}$ of the form

$$\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

A Wald test statistic associated with the null $H_0^{(t)}$ can thus be constructed as

$$\{\mathbf{C}^{(t)} \mathrm{vec}(\hat{\boldsymbol{\gamma}}_{jl}^\top)\}^\top \left[\mathbf{C}^{(t)} \widehat{\mathbf{M}}_{jl} \{\mathbf{C}^{(t)}\}^\top\right]^{-1} \mathbf{C}^{(t)} \mathrm{vec}(\hat{\boldsymbol{\gamma}}_{jl}^\top), \tag{4.5}$$

where $\widehat{\mathbf{M}}_{jl}$ denotes the lth $K\breve{K} \times K\breve{K}$ diagonal block of $\{-\sum_{g=1}^{G} \ddot{\ell}_{jg}(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\theta}}_j)\}^{-1}$ with $\ddot{\ell}_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ being the Hessian matrix of $\ell_{jg}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$. Under $H_0^{(t)}$, the test statistic approximately follows a chi-squared distribution with $\breve{K}(K-1)$ degrees of freedom.

To test whether $\beta_{jl}(t, \breve{x})$ varies significantly with $\breve{x}$, observe that when $\gamma_{jl\breve{k}k} \equiv \gamma_{jl\cdot k}$ for any $k = 1, \ldots, K$, $\breve{k} = 1, \ldots, \breve{K}$, the bivariate coefficient $\beta_{jl}(t, \breve{x}) = \sum_{k=1}^{K} \gamma_{jl\cdot k} B_k(t)$ no longer varies with $\breve{x}$. The corresponding null hypothesis is $H_0^{(\breve{x})} : \mathbf{C}^{(\breve{x})} \mathrm{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$

where $\mathbf{C}^{(\breve{x})}$ is a $K(\breve{K}-1)\times K\breve{K}$ difference matrix of the Kth order. The Wald test statistic is readily obtained by substituting $\mathbf{C}^{(t)}$ in (4.5) with $\mathbf{C}^{(\breve{x})}$. Similarly, the null hypothesis for testing whether $\beta_{jl}(t,\breve{x})$ varies significantly with both t and $\breve{x}$ is $H_0^{(t,\breve{x})}:\mathbf{C}^{(t,\breve{x})}\mathrm{vec}(\boldsymbol{\gamma}_{jl}^\top)=\mathbf{0}$, where $\mathbf{C}^{(t,\breve{x})}$ is a $(K\breve{K}-1)\times K\breve{K}$ first-order difference matrix. The Wald test statistic can be written by substituting $\mathbf{C}^{(t)}$ in (4.5) with $\mathbf{C}^{(t,\breve{x})}$.

4.4 Penalized partial likelihood approach

4.4.1 Difference-based anisotropic penalization

To mitigate overfitting and increase the smoothness of the estimated coefficient surface of $\beta_{jl}(\cdot,\cdot)$, we consider penalizing the column-wise and row-wise differences between adjacent control points of $\boldsymbol{\gamma}_{jl}$. The penalized log-partial likelihood can be written as

$$\ell^{(\mathrm{P})}(\boldsymbol{\gamma},\boldsymbol{\theta})=\sum_{j=1}^{m}\ell_j^{(\mathrm{P})}(\boldsymbol{\gamma}_j,\boldsymbol{\theta}_j;\boldsymbol{\mu}_j,\breve{\boldsymbol{\mu}}_j),$$

where

$$\begin{aligned}\ell_j^{(\mathrm{P})}(\boldsymbol{\gamma}_j,\boldsymbol{\theta}_j;\boldsymbol{\mu}_j,\breve{\boldsymbol{\mu}}_j)&\coloneqq\ell_j(\boldsymbol{\gamma}_j,\boldsymbol{\theta}_j)-\sum_{l=1}^{p}\left\{\breve{\mu}_{jl}^2\|\breve{\mathbf{D}}\boldsymbol{\gamma}_{jl}\|_{\mathrm{F}}^2+\mu_{jl}^2\|\boldsymbol{\gamma}_{jl}\mathbf{D}^\top\|_{\mathrm{F}}^2\right\}\\&=\ell_j(\boldsymbol{\gamma}_j,\boldsymbol{\theta}_j)-\boldsymbol{\gamma}_j^\top\mathbf{P}_j(\boldsymbol{\mu}_j,\breve{\boldsymbol{\mu}}_j)\boldsymbol{\gamma}_j.\end{aligned}\tag{4.6}$$

In (4.6), $\boldsymbol{\mu}_j\coloneqq[\mu_{j1},\mu_{j2},\ldots,\mu_{jp}]^\top$ and $\breve{\boldsymbol{\mu}}_j\coloneqq[\breve{\mu}_{j1},\breve{\mu}_{j2},\ldots,\breve{\mu}_{jp}]^\top$) denote vectors of smoothing parameters controlling the amount of penalty, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, and

$$\begin{aligned}\mathbf{P}_j(\boldsymbol{\mu}_j,\breve{\boldsymbol{\mu}}_j)\coloneqq&\{\breve{\mathbf{D}}\otimes\mathbf{I}\otimes\mathrm{diag}(\breve{\boldsymbol{\mu}}_j)\}^\top\{\breve{\mathbf{D}}\otimes\mathbf{I}\otimes\mathrm{diag}(\breve{\boldsymbol{\mu}}_j)\}\\&+\{\breve{\mathbf{I}}\otimes\mathbf{D}\otimes\mathrm{diag}(\boldsymbol{\mu}_j)\}^\top\{\breve{\mathbf{I}}\otimes\mathbf{D}\otimes\mathrm{diag}(\boldsymbol{\mu}_j)\},\end{aligned}$$

where $\mathbf{I}$ (or $\breve{\mathbf{I}}$) is a $K\times K$ (or $\breve{K}\times\breve{K}$) identity matrix, $\mathrm{diag}(\cdot)$ converts a vector into a diagonal matrix, and $\breve{\mathbf{D}}$ (or $\mathbf{D}$) is a $(\breve{K}-1)\times\breve{K}$ (or $(K-1)\times K$) first-order

difference matrix. In what follows, we use tildes to indicate penalized estimates. For example, $\tilde{\boldsymbol{\gamma}}_{jl}$ denotes a vector of penalized estimates of $\boldsymbol{\gamma}_{jl}$.

4.4.2 Asymptotics

Before presenting the inferential procedures with penalization, we first derive the asymptotic distribution of the penalized estimates $\tilde{\boldsymbol{\eta}}_j \coloneqq [\tilde{\boldsymbol{\gamma}}_j^\top, \tilde{\boldsymbol{\theta}}_j^\top]^\top$. We assume that the knot locations, K, $\breve{K}$, p, and q remain fixed as the sample size $n \coloneqq \sum_{g=1}^{G} n_g$ increases. Observe that as n grows, the contribution from the log-partial likelihood in (4.6) increases. To preserve the degree of smoothness, $\boldsymbol{\mu}_j$ and $\breve{\boldsymbol{\mu}}_j$ will need to increase at a rate of $O(\sqrt{n})$. Here we consider two cases when the contribution of the penalty term to the penalized score function, $\mathbf{P}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\boldsymbol{\gamma}_j$, is not necessarily $\mathbf{0}$, but the amount of smoothing (and the introduced bias) shrinks as n increases. First, given two constants $\mu_{jl}^{(0)}$ and $\breve{\mu}_{jl}^{(0)}$, if $\mu_{jl}/n^{1/4} \to \mu_{jl}^{(0)}$ and $\breve{\mu}_{jl}/n^{1/4} \to \breve{\mu}_{jl}^{(0)}$ as n increases [45], then standard derivations imply that $\sqrt{n}(\tilde{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)$ is asymptotically normal with a mean estimate

$$\sqrt{n}\tilde{\mathbf{b}}_j \coloneqq \sqrt{n}\left\{\ddot{\ell}_j^{(\mathrm{P})}(\tilde{\boldsymbol{\eta}}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\right\}^{-1} \mathbf{Q}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\tilde{\boldsymbol{\eta}}_j$$

and a sandwich estimate of variance

$$n\widetilde{\mathbf{V}}_j^{\mathrm{S}} \coloneqq -n\left\{\ddot{\ell}_j^{(\mathrm{P})}(\tilde{\boldsymbol{\eta}}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\right\}^{-1} \ddot{\ell}_j(\tilde{\boldsymbol{\eta}}_j) \left\{\ddot{\ell}_j^{(\mathrm{P})}(\tilde{\boldsymbol{\eta}}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\right\}^{-1},$$

where

$$\ddot{\ell}_j^{(\mathrm{P})}(\boldsymbol{\eta}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) = \ddot{\ell}_j(\boldsymbol{\eta}_j) - \mathbf{Q}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)$$

is the penalized Hessian matrix of (4.6), and $\mathbf{Q}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)$ is a block diagonal matrix with two blocks $\mathbf{P}_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)$ and $\mathbf{0}$ (a $q \times q$ matrix). As a second case, if $\mu_{jl}/n^{1/4} \to 0$ and $\breve{\mu}_{jl}/n^{1/4} \to 0$ as n increases, the variance of $(\tilde{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)$ can be well approximated by the inverse of the penalized information matrix, i.e., $\widetilde{\mathbf{V}}_j^{\mathrm{M}} = -\left\{\ddot{\ell}_j^{(\mathrm{P})}(\tilde{\boldsymbol{\eta}}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\right\}^{-1}$, a model-based variance estimate.

4.4.3 Inference

In the presence of penalization, a Wald test statistic associated with the null hypothesis $H_0^{(t)} : \mathbf{C}^{(t)}\text{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$ can be written as

$$\{\text{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}^\top \{\mathbf{C}^{(t)}\}^\top \left[\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top\right]^{-1} \mathbf{C}^{(t)}\{\text{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}, \tag{4.7}$$

where $\tilde{\mathbf{b}}_{jl}$ denotes the lth $K\breve{K}$-dimensional subvector of $\tilde{\mathbf{b}}_j$, and $\boldsymbol{\Omega}_{jl}$ denotes an arbitrary $K\breve{K} \times K\breve{K}$ symmetric and positive-definite matrix, e.g., the lth diagonal block of $\widetilde{\mathbf{V}}_j^{\text{S}}$ or $\widetilde{\mathbf{V}}_j^{\text{M}}$. The distribution of the test statistic (4.7) is characterized in Proposition IV.1 below. The proof is available in Appendix C.2.

Proposition IV.1. *Under $H_0^{(t)}$, the test statistic* (4.7) *asymptotically follows a distribution characterized by*

$$\sum_{u=1}^{K\breve{K}\times K\breve{K}} \mu_u G_u^2,$$

where G_u's are independent standard normal random variables, and μ_u's are the possibly identical eigenvalues of the matrix product of $[\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top]^{-1}$ and the variance of $\mathbf{C}^{(t)}\{\text{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$.

Similarly as in Section 4.3.2, for the null $H_0^{(\breve{x})} : \mathbf{C}^{(\breve{x})}\text{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$, the corresponding Wald test statistic can be obtained by substituting $\mathbf{C}^{(t)}$ in (4.7) with $\mathbf{C}^{(\breve{x})}$. For the null $H_0^{(\breve{x})} : \mathbf{C}^{(t,\breve{x})}\text{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$, the Wald test statistic can be written by substituting $\mathbf{C}^{(t)}$ in (4.7) with $\mathbf{C}^{(t,\breve{x})}$.

4.4.4 Cross-validated parameter tuning

To identify an optimal set of tuning parameters to alleviate model overfitting and the unsmoothness of the estimated effect surface, we consider 5 methods of cross-validation. In the first 4 methods, the entire data sample needs to be partitioned into F subsamples (hereafter folds) of approximately equal sizes. For failure type j

and $f = 1, \ldots, F$, let $\tilde{\boldsymbol{\eta}}_j^{-f}$ be the penalized estimates of $\boldsymbol{\eta}_j$ based on the complement of fold f, and let ℓ_j^f and ℓ_j^{-f} be the (unpenalized) log-partial likelihood based on fold f and the complement of fold f, respectively. A cross-validation error (CVE) for failure type j is then defined in each of the 4 approaches. The last method of generalized cross-validation does not require data partitioning in the calculation of CVE. Optimal tuning parameters can be determined through minimizing the CVE. A comprehensive evaluation of the 5 approaches is presented in Section 4.5.2.

Fold-constrained (FC) cross-validated partial likelihood

In this approach, the CVE is proportional to the sum of fold-specific log-partial likelihood functions in which risk sets are constrained by the corresponding folds, i.e.,

$$\text{CVE}_j \coloneqq -2 \sum_{f=1}^{F} \ell_j^f(\tilde{\boldsymbol{\eta}}_j^{-f}).$$

Complementary fold-constrained (CFC) cross-validated partial likelihood

As the name suggests, the CVE is proportional to the sum of complementary fold-constrained log-partial likelihood functions, i.e.,

$$\text{CVE}_j \coloneqq -2 \sum_{f=1}^{F} \{\ell_j(\tilde{\boldsymbol{\eta}}_j^{-f}) - \ell_j^{-f}(\tilde{\boldsymbol{\eta}}_j^{-f})\}.$$

This approach was applied in Verweij and Van Houwelingen (1993) [122] and Simon et al. (2011) [110].

Unconstrained (UC) cross-validated partial likelihood

First introduced by Breheny and Huang (2011) [10], this approach features risk set construction unconstrained by folds in that fold-specific estimates $\tilde{\boldsymbol{\eta}}_j^{-f}$'s are assigned to all units of the sample according to their fold identities. With a slight

abuse of notation, the CVE is written as

$$\text{CVE}_j \coloneqq -2\ell_j(\tilde{\boldsymbol{\eta}}_j^{-1}, \ldots, \tilde{\boldsymbol{\eta}}_j^{-F}),$$

where $\tilde{\boldsymbol{\eta}}_j^{-f}$ is assigned to observations of fold f.

Cross-validated deviance residuals (DR)

Dai and Breheny (2019) [22] used the sum of squared deviance residuals [119] as a criterion of cross-validation in a penalized Cox proportional hazards model. However, their approach cannot be directly applied to a non-proportional hazards model with varying coefficients. To proceed, we first derive the deviance residuals for model (4.1) in the next proposition, the proof of which is available in Appendix C.3.

Proposition IV.2. *Let $\hat{\lambda}_{0jg}(\cdot)$ be the estimated baseline hazard function derived from the unpenalized bivariate varying coefficient model. Let*

$$\tilde{M}_{jgi} \coloneqq \Delta_{jgi} - \exp(\mathbf{W}_{gi}^\top \tilde{\boldsymbol{\theta}}_j^{-f}) \int_0^{X_{gi}} \exp\left\{\mathbf{Z}_{gi}^\top \tilde{\boldsymbol{\beta}}_j^{-f}(t, \breve{X}_{gi})\right\} \hat{\lambda}_{0jg}(t)\,\mathrm{d}t$$

be the martingale residual for subject i in the gth stratum, where $\tilde{\boldsymbol{\beta}}_j^{-f}(\cdot,\cdot)$ and $\tilde{\boldsymbol{\theta}}_j^{-f}$ are the penalized estimates from the corresponding fold f to which subject i in the gth stratum belongs. Then the deviance residual for subject i in the gth stratum with respect to the jth failure type is written as

$$d_{jgi} \coloneqq \text{sign}(\tilde{M}_{jgi})\sqrt{-2\left[\Delta_{jgi}\left\{\mathbf{Z}_{gi}^\top \tilde{\boldsymbol{\beta}}_j^{-f}(X_{gi}, \breve{X}_{gi}) + \mathbf{W}_{gi}^\top \tilde{\boldsymbol{\theta}}_j^{-f} + \log \int_0^{X_{gi}} \hat{\lambda}_{0jg}(t)\,\mathrm{d}t\right\} + \tilde{M}_{jgi}\right]}.$$

Given the deviance residuals in Proposition IV.2, the CVE can be written as

$$\text{CVE}_j \coloneqq \sum_{g=1}^{G}\sum_{i=1}^{n_g} d_{jgi}^2.$$

Generalized cross-validation (GCV)

Extending the approach of Yan and Huang (2012) [140] to this setting with bivariate varying coefficients, we can write the CVE for the jth failure type as

$$\text{CVE}_j = -\frac{\ell_j(\boldsymbol{\eta}_j)}{n(1 - f_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)/n)^2},$$

where $f_j(\boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j) \coloneqq \text{trace}\left(\{\ddot{\ell}_j^{(\text{P})}(\boldsymbol{\eta}_j; \boldsymbol{\mu}_j, \breve{\boldsymbol{\mu}}_j)\}^{-1}\ddot{\ell}_j(\boldsymbol{\eta}_j)\right)$, i.e., the number of effective parameters [140], or the "degrees of freedom" of the model [45].

4.5 Simulation experiments

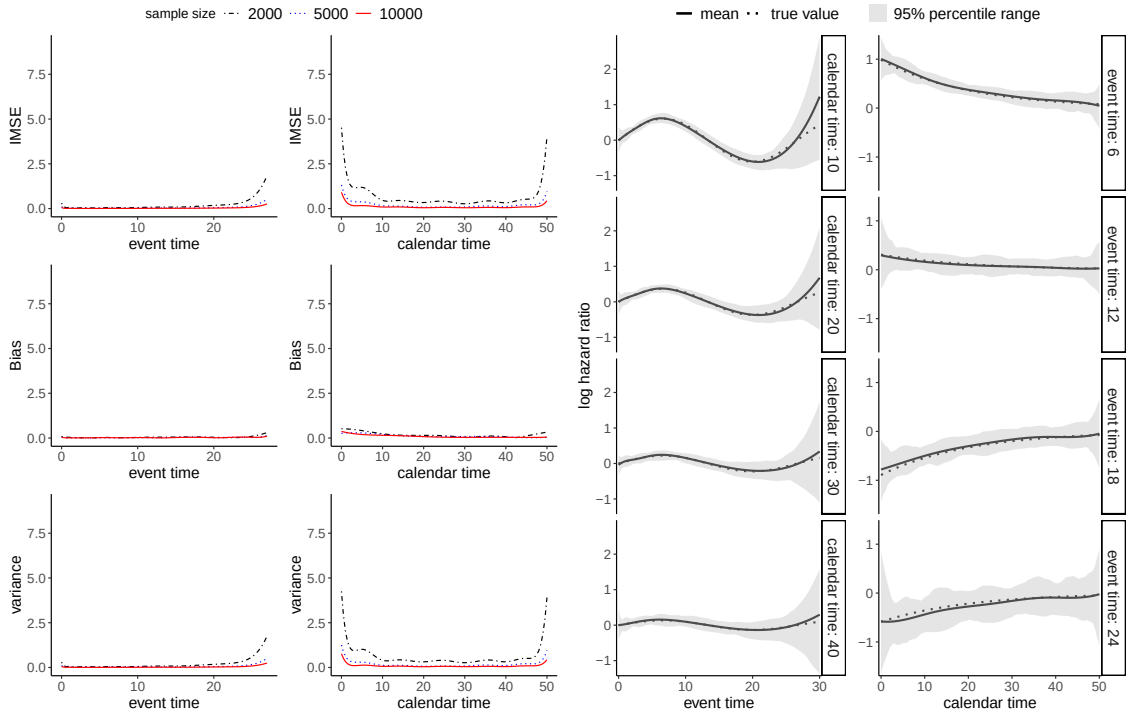
4.5.1 Unpenalized approach

Following the approach in Section 4.3, we assessed the bivariate varying coefficient model for competing risks via simulation experiments. Since distinct types of competing risks can be analyzed separately within a cause-specific hazard framework, we focused on a single event type and dropped the subscript j to allow simplified notation. Therefore, no stratification was used in the data generating process.

In each simulation scenario, a number (100 or 1,000) of independent data replicates were generated with the sample size varying from 1,000 to 10,000. For each sample unit, two covariates (corresponding to $\mathbf{Z}_{gi}$ and $\mathbf{W}_{gi}$ in Section 4.2) were drawn from a bivariate normal distribution with zero mean, one variance, and correlation $\rho = 0.6$. Two coefficients were set as $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$, with event time t varying from 0 to 30, and calendar time $\breve{x}$ varying from 0 to 50. Underlying event times were determined via a root-finding procedure based on the cause-specific hazard function in Section 4.2 [7]. Calendar times were drawn from a uniform distribution bounded by 0 and 50. Censoring times were sampled from a uniform distribution bounded by 0 and 30. Observed event times were determined as the minimum of the underlying event and censoring time pairs.

Figure 4.2a presents the integrated mean squared error (IMSE), bias, and variance (all averaged over a grid of 100 evenly spaced points across either event or calendar time) with respect to the bivariate varying coefficient $\beta_1(t, \breve{x})$ with the sample size growing from 2,000 to 10,000. The three metrics were calculated based on 100 data replicates. The coefficient β_2 was treated as a time-invariant parameter in model fitting. On the event timescale, the IMSE becomes higher as event time increases, due to the fact that the shrinking risk set leads to fewer remaining units in the sample and hence less accurate estimation. As the sample size grows, the IMSE curve shifts downward and the IMSE is substantially reduced towards the end of follow-up. On the calendar timescale, the IMSE is higher on both ends and the curve becomes lower as the sample size increases from 2,000 to 10,000. Moreover, a comparison between the second and third row of Figure 4.2a suggests that the IMSE on both timescales is predominantly determined by the variance component. As a complement to Figure 4.2a, Figure 4.2b provides additional evidence on estimation, with the sample size fixed at 10,000. Throughout all panels of distinct event and calendar times, the mean estimated curve tracks closely with the true effect curve, demonstrating the accurate estimation of the extended proximal Newton algorithm.

At different event and calendar times, we compared coverage probability (CP) curves of $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ with varied sample sizes in Figure 4.3a. Pointwise 95% confidence intervals were used throughout all panels. Overall, all curves remained around the 0.95 reference line, except that the CP dropped below 0.8 toward the end of the event time period with calendar time equal to 10 and sample size equal to 10,000. In Figure 4.3b, we evaluated three tests of univariate and bivariate variation with respect to $\beta_1(t, \breve{x})$, where the sample size varied from 2,000 to 10,000. As expected [136], curves of type I error rate were sloping downward



(a) IMSE, bias, and variance

(b) Mean versus true value

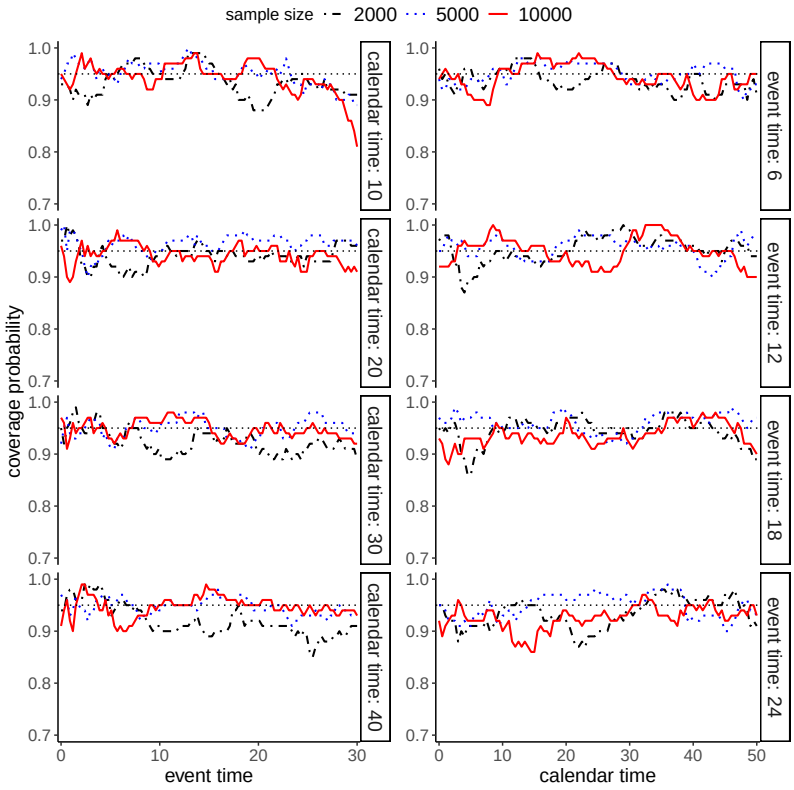
Figure 4.2: (a) Integrated mean squared error (IMSE), average bias, and average variance of the estimated surface $\hat{\beta}_1(t, \breve{x})$ with varied sample sizes on event and calendar timescales. In each scenario, 100 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. (b) Mean and 95% percentile range (2.5th and 97.5th percentiles as lower and upper limits) of pointwise estimates of $\beta_1(t, \breve{x})$ at selected event times and calendar times. In each scenario, 100 data replicates were generated with sample size equal to 10,000. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. An unpenalized approach was used in (a) and (b).

with the sample size, while the rates remained slightly higher than 0.05 as the sample size exceeded 5,000. The power grew dramatically until the sample size reached 4,000, and then remained higher than 0.95 afterward.

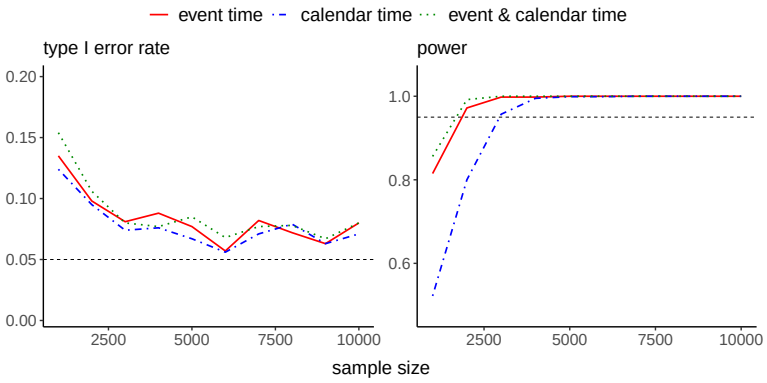
4.5.2 Penalized approach

In similar simulation settings, we evaluated the difference-based anisotropic penalization and corresponding tests of effect variation. With sample size fixed at $n = 10{,}000$, Figure 4.4a shows the IMSE, bias, and variance (again, averaged over a grid a 100 evenly spaced points on either time scale) of the estimated effect surface $\hat{\beta}_1(t, \breve{x})$ on two timescales with different pairs of tuning parameters μ and $\breve{\mu}$, where the unpenalized approach with $\mu = 0$ and $\breve{\mu} = 0$ is included as a reference. Across both event and calendar time, the IMSE was the highest when the penalty was minimal ($\mu = 0.02$ and $\breve{\mu} = 0.05$). As the penalty became more prominent ($\mu = 0.2$ and $\breve{\mu} = 0.5$), the IMSE decreased at first, especially on the calendar timescale. When $\mu = 2$ and $\breve{\mu} = 5$, the IMSE rebounded substantially. As for bias, a higher penalty level led to a higher bias across event time, while the bias remained lowest with $\mu = 0.2$ and $\breve{\mu} = 0.5$. Unsurprisingly, a higher level penalty was associated with lower variance for both timescales. This result suggests that $\mu = 0.2$ and $\breve{\mu} = 0.5$ are the optimal pair among the four. This pair was applied exclusively in Figure 4.4b, where the curves of true values were compared to the curves of mean estimates. In all panels, the two curves tracked closely, except toward the end of the event time.

Figure 4.5 presents the CP, type I error rate, and power with varying sample sizes. To allow tuning parameters to vary with sample size, we set $\mu = 0.002n^{1/8}$ and $\breve{\mu} = 0.005n^{1/8}$, corresponding to the second case in Section 4.4.2. Across event and calendar time, the CP curve fluctuated closely around the 0.95 reference line, except that the CP dropped to 0.75 toward the end of the event time period with calen-

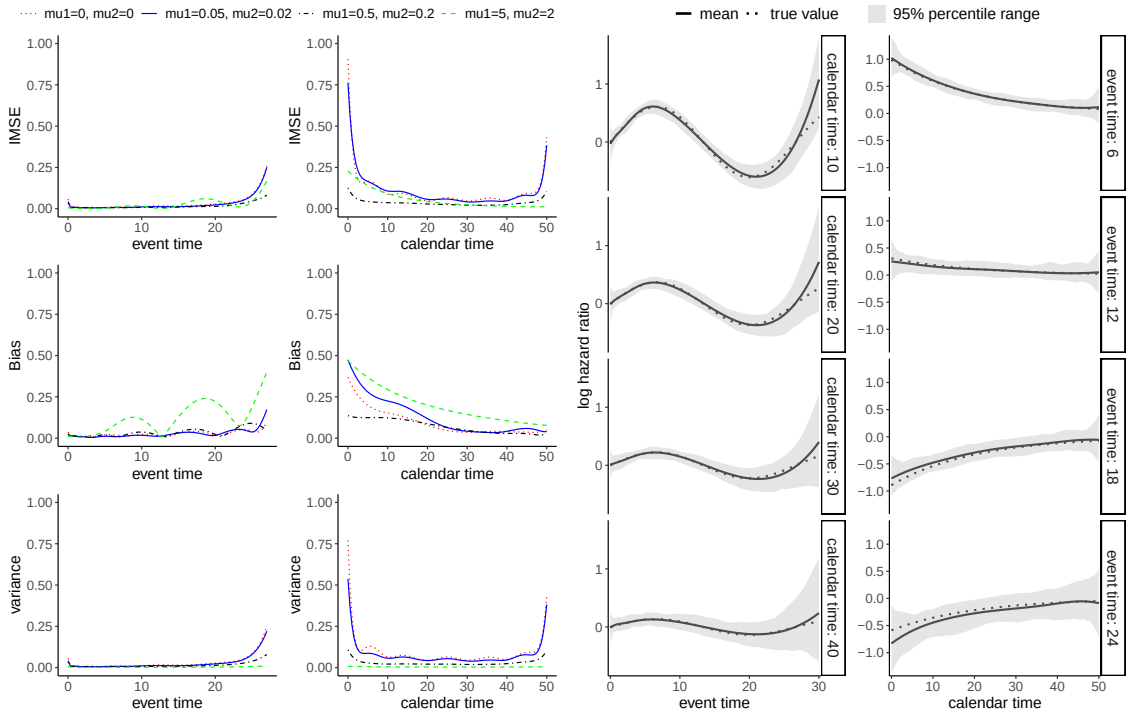


(a) Coverage probability



(b) Type I error rate and power

Figure 4.3: (a) Coverage probability curves of $\beta_1(t, \breve{x})$ via pointwise 95% confidence intervals on event and calendar time scales, with varied sample sizes. In each scenario, 100 data replicates were generated with sample size equal to 10,000. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. (b) Type I error rate and power curves for tests of univariate and bivariate variation with varied sample sizes. In each scenario, 1,000 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = 1$ and $\beta_2 = 1$ in the left panel, and $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$ in the right panel. An unpenalized approach was used in (a) and (b).



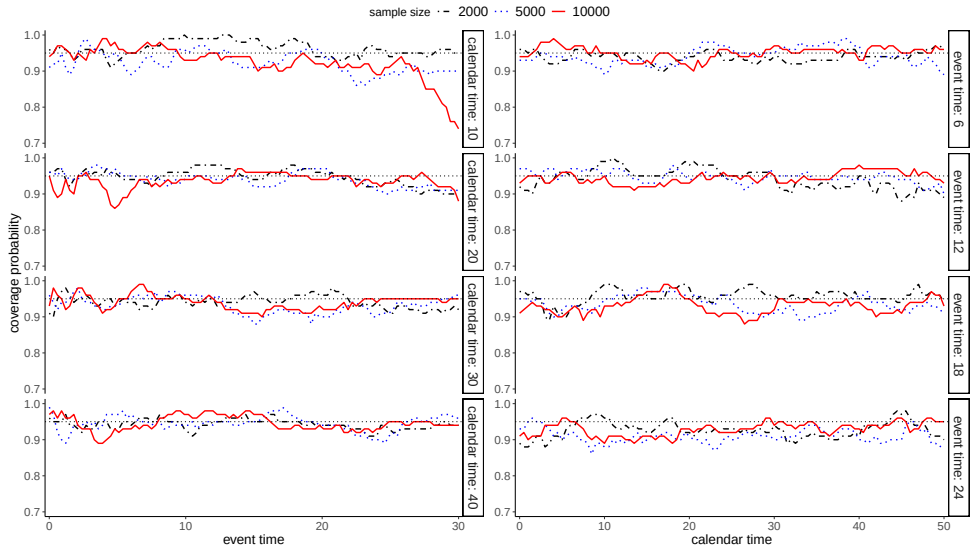
(a) IMSE, bias, and variance

(b) Mean versus true value

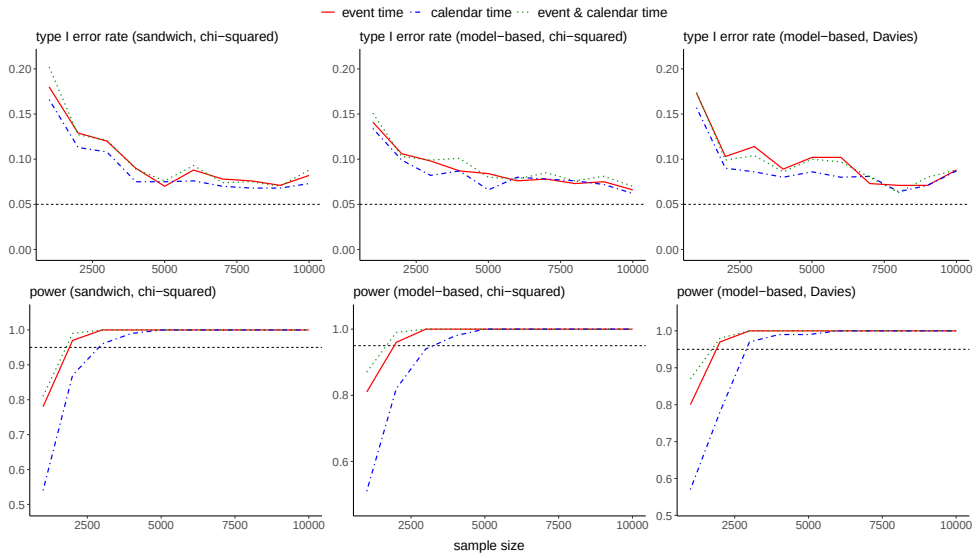
Figure 4.4: (a) Integrated mean squared error (IMSE), average bias, and average variance of the estimated surface $\hat{\beta}_1(t, \breve{x})$ with sample size fixed at 10,000. In each scenario, 100 data replicates were generated. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. Various levels of penalization were introduced to $\beta_1(\cdot, \cdot)$, where mu1 and mu2 denote tuning parameters for calendar and event time, respectively, as in (4.6). (b) Mean and 95% percentile range (2.5th and 97.5th percentiles as lower and upper limits) of pointwise estimates of $\beta_1(t, \breve{x})$ at selected event times and calendar times. In each scenario, 100 data replicates were generated with sample size equal to 10,000. On both timescales, $K = \breve{K} = 7$ cubic ($d = \breve{d} = 3$) B-spline functions form a basis. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. Only the optimal case in Part (a), i.e., mu1=0.5 and mu2=0.2, was considered.

dar time equal to 10 and sample size equal 10,000 (Figure 4.5a, top left panel). In each column of Figure 4.5b, we adopted a distinct construction of the test statistics based on (4.7), and considered three tests of variation, jointly and separately. In the first and second columns, the sandwich and model-based variance estimators, respectively, were employed to determine $\boldsymbol{\Omega}$ and the variance of $\text{vec}(\tilde{\boldsymbol{\gamma}}^\top) - \tilde{\mathbf{b}}$, respectively, so that the test statistics approximately followed the chi-squared distribution. In the third column, the model-based estimate was used to form $\boldsymbol{\Omega}$, while the variance of $\text{vec}(\tilde{\boldsymbol{\gamma}}^\top) - \tilde{\mathbf{b}}$ was estimated via the sandwich estimator. The resulting test statistics, similar to the one in Gray (1992) [45], approximately followed a distribution characterized by a linear combination of chi-squared random variables [23]. This distribution was implemented via the package CompQuadForm [66]. We observed that the third construction generally led to higher type I error rates than the other two. When sample size was up to 3,000, the model-based test statistics gave lower type I error rates; when sample size exceeded 3,000, the sandwich test statistics overall resulted in slightly lower type I error rates. All three constructions were associated with sufficiently high power with sample size greater than or equal to 3,000.

To compare the five methods of cross-validation via simulations in Section 4.4.4, we generated 100 pairs of training and testing data replicates for each sample size n (varying from 2,000 to 5,000). A 5-by-5 grid of tuning parameters was formed such that $\mu/\sqrt{n}$ and $\breve{\mu}/\sqrt{n}$ varied from 10^{-5} to 10^{-1}. All five methods were applied to a training copy to obtain an optimal pair of tuning parameters and penalized estimates. The training data were split into four folds whenever data partitioning was necessary. The penalized estimates were then applied to both training and testing replicates in the calculation of -2ℓ (ℓ denoting the unpenalized log partial likelihood) and the average IMSE, two measures of predictive accuracy used for eval-



(a) Coverage probability



(b) Type I error rate and power

Figure 4.5: (a) Coverage probability curves of $\beta_1(t, \breve{x})$ via pointwise 95% confidence intervals at varied event time, calendar time, and sample sizes. In each scenario, 100 data replicates were generated with sample size $n = 10,000$. True values are $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. (b) Type I error rate and power curves for tests of univariate and bivariate variation with different test statistics and varied sample sizes. In each scenario, 1,000 data replicates were generated. True values are $\beta_1(t, \breve{x}) = 1$ and $\beta_2 = 1$ in the top 3 panels, and $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$ in the bottom 3 panels. In the first and second column, a sandwich and a model-based variance estimator were used with test statistics approximately following a chi-squared distribution. In the third column, the test statistic in Gray (1992)[45] was compared with a distribution of a linear combination of chi-squared random variables [23]. In Parts (a) and (b), 7 cubic B-splines form a basis on both timescales, and tuning parameters vary with sample size, i.e., $\mu = n^{1/8}/500$ and $\breve{\mu} = n^{1/8}/200$.

uating the five methods. The distribution of selected tuning parameters is reported in the Web Appendix D of the Supporting Information, and -2ℓ and average IMSE for all methods are tabulated in Table 4.1. The method of cross-validated deviance residuals (DR) led to the lowest -2ℓ when the sample size of the training data was less than 5,000, or when the sample size of the testing data was 3,000 or 5,000; it also led to the lowest average IMSE when the sample size of the training data was 4,000 or 5,000. In contrast, the generalized cross-validation was associated with the highest -2ℓ for both training and testing data across different sample sizes; it also gave the highest average IMSE except when the sample size was 2,000. Although DR overall achieved the highest predictive accuracy, its advantage over the other 3 data-partitioning cross-validation methods was not significant.

Table 4.1: A simulation-based comparison of five cross-validation methods: fold-constrained (FC), complementary fold-constrained (CFC), and fold-unconstrained (UC) cross-validated partial likelihood, cross-validated deviance residuals (DR), and generalized cross-validation (GCV). In each scenario, 100 training and validation data replicates were generated independently. Each cross-validation method was applied to the training data replicate to obtain the penalized estimates. The estimates were then applied to the training and validation data separately to calculate -2ℓ (Panel A), where ℓ denotes the unpenalized log partial likelihood, and to the training data to calculate average integrated mean squared error (IMSE, Panel B). For IMSE, the average was taken across 10,201 different combinations of event and calendar time. True values were $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2 = 1$. Standard deviations are provided in parentheses.

Panel A: -2ℓ

sample size	training					testing				
	FC	CFC	UC	DR	GCV	FC	CFC	UC	DR	GCV
2000	12951.11 (2636.17)	12950.53 (2634.96)	12951.02 (2634.98)	12949.77 (2635.09)	12978.45 (2635.43)	11917.90 (11.80)	11917.24 (10.72)	11917.69 (10.90)	11917.68 (10.85)	11927.87 (4.70)
3000	20729.73 (4174.25)	20729.64 (4174.46)	20729.66 (4174.56)	20729.58 (4174.55)	20769.08 (4176.98)	25852.29 (16.02)	25852.16 (16.01)	25852.21 (15.99)	25852.12 (15.98)	25871.91 (8.77)
4000	28696.21 (5753.25)	28696.33 (5752.60)	28696.21 (5752.59)	28695.57 (5753.07)	28746.12 (5757.26)	26794.12 (10.51)	26794.10 (10.49)	26794.11 (10.50)	26794.34 (10.57)	26831.69 (10.30)
5000	37036.32 (7439.08)	37035.18 (7439.23)	37035.27 (7439.38)	37036.26 (7439.86)	37095.90 (7451.81)	54238.77 (26.10)	54239.27 (26.37)	54239.44 (26.19)	54238.45 (26.85)	54279.14 (14.08)

Panel B: average IMSE

sample size	training				
	FC	CFC	UC	DR	GCV
2000	0.1023 (0.1193)	0.1019 (0.1137)	0.1041 (0.1148)	0.1104 (0.1170)	0.0812 (0.0124)
3000	0.0697 (0.0679)	0.0710 (0.0684)	0.0706 (0.0685)	0.0726 (0.0678)	0.0775 (0.0100)
4000	0.0714 (0.0917)	0.0703 (0.0917)	0.0680 (0.0904)	0.0674 (0.0917)	0.0747 (0.0098)
5000	0.0562 (0.1193)	0.0567 (0.1137)	0.0570 (0.1148)	0.0550 (0.1170)	0.0729 (0.0124)

4.6 Applications to dialysis patients amidst COVID-19

To better understand the dynamics of the COVID-19 effect on dialysis patients, we applied the bivariate varying coefficient model to two large-scale retrospective studies, both having data abstracted from the CMS clinical and administrative database (primarily based on the Renal Management Information System, CROWNWeb facility-reported clinical and administrative data, the Medicare Enrollment Database, and Medicare claims data). In both studies, the interest was in the impact of an in-hospital COVID-19 diagnosis on the outcomes of dialysis patients. Information on in-hospital COVID-19 diagnosis was mainly obtained from Medicare inpatient and physician/supplier claims. An in-hospital COVID-19 diagnosis was confirmed if the patient's inpatient or physician/supplier claim associated with the hospitalization had either of the two diagnosis codes of the International Classification of Diseases, 10th Revision: B97.29 or U07.1 [134]. In addition to COVID-19, a comprehensive list of patient demographics, clinical characteristics, and prevalent comorbidities were considered as baseline risk factors.

4.6.1 Postdischarge outcomes

In the first study, outcomes of primary interest were all-cause unplanned acute-care-hospital readmission and death within 30 days of hospital discharge. This study consisted of 436,745 live acute-care hospital discharges of 222,154 Medicare beneficiaries on dialysis from 7,871 Medicare-certified dialysis facilities between January 1, 2020 and October 31, 2020. Discharges from non-acute care hospitals, discharges with in-hospital death, and discharges with discharge-day outcomes were excluded from the data, along with other administrative exclusions.

The 8 panels of Figure 4.6 show different perspectives of the bivariate dynam-

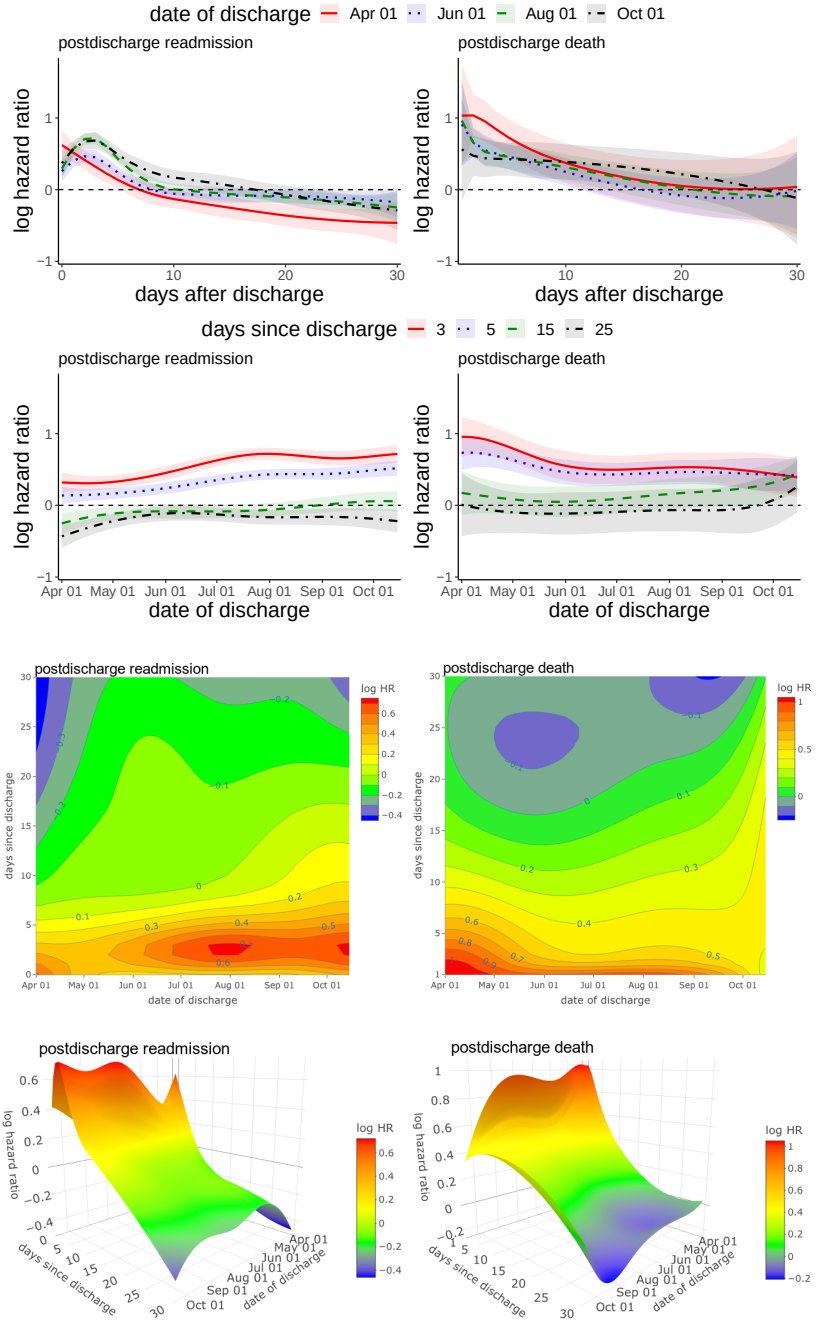


Figure 4.6: Bivariate variation of log hazard ratios with respect to in-hospital COVID-19 diagnosis for 30-day postdischarge readmission and death. Included in the sample were 436,745 live hospital discharges of 222,154 Medicare beneficiaries on dialysis from 7,871 Medicare-certified dialysis facilities from January 1, 2020 to October 31, 2020. Ribbons in the top four panels indicate 95% confidence intervals. Panels in the third and fourth rows are contour and surface plots, respectively.

ics of the COVID-19 effect in terms of the log hazard ratio on 30-day postdischarge readmission and death. The penalized likelihood approach was used to improve the smoothness of the estimated surface, with tuning parameters determined by the method of cross-validated deviance residuals. The two panels in the first row present 30-day postdischarge variations at 4 distinct dates of discharge. The downward sloping curves indicate that having COVID-19 was associated with significantly elevated risks of readmission and death, but only over the first week of discharge. The two panels in the second row present variations with calendar time on 4 different days after discharge, where the COVID-19 effect became less significant with more days since discharge. Within the first 5 days of discharge, the risk of readmission gradually increased as the pandemic unfolded, whereas the risk of death decreased until early June and then remained relatively unchanged afterward.

The remaining panels in the third and fourth rows of Figure 4.6 are contour and surface plots, respectively, displaying the variations of the COVID-19-associated risks of readmission and death along two dimensions of time. Persistently declining log hazard ratios were observed from Day 0 to Day 30 since discharge, suggesting that the COVID-19 effects on readmission and death were decreasing with time. During the first 5 days after discharge, there existed three peaks for readmissions around early April, early August, and mid October of 2020, while there was only one peak for death in early April. These findings are consistent with the evolution of the COVID-19 pandemic in the general population. In the initial phase of the pandemic, the case fatality rate was extremely high as the highly pathogenic variants of the novel coronavirus hit the country. Restricted access to health services and the fear of contagion contributed to deferred hospitalizations and readmissions, which supports the mildly high risk of readmission in early April. As governments implemented

various mandates to contain the spread of the coronavirus, patients became more willing to be admitted to hospital, with the risk of readmission rebounding. In the meantime, the pervasive variants were getting less pathogenic, and hospitals became more prepared to treat COVID-19 patients, both of which led to a reduced case fatality rate. The risk of postdischarge death therefore decreased with calendar time.

In addition to modeling the bivariate COVID-19 effect on postdischarge outcomes, we tested its variation along two time dimensions according to Section 4.4.3. Consistent with the top right panel of Figure 4.6, the test of univariate variation across calendar time for postdischarge death led to a p-value of 0.727, indicating that the risk of death did not vary significantly with calendar time. All other tests of univariate and bivariate variation led to p-values less than 0.001.

4.6.2 Discharge destinations

In the second study, outcomes of interest were three options of discharge destination, including (1) in-hospital death or discharge to hospice, (2) discharge to a long- or short-term care hospital, skilled nursing facility, intermediate care facility, inpatient rehabilitation facility, psychiatric hospital, or critical access hospital (hereafter discharge to another facility), and (3) discharge to home with or without home care services, together viewed as mutually exclusive competing risks. Included in the data were 544,677 unplanned hospitalizations of 250,940 Medicare dialysis beneficiaries associated with 2,929 dialysis facilities throughout the year of 2020 (determined based on admission dates). Each hospital admission was followed up for up to 40 days. Among the 544,677 hospital admissions, 44,858 resulted in an in-hospital death or discharge to hospice; 125,723 were followed by a discharge to another facility; and 371,104 resulted in a home discharge. The remaining 2,992 admissions were

associated with a hospital stay longer than 40 days, i.e., a censoring.

We ran an unpenalized bivariate varying coefficient model to validate its performance on the discharge status data, in which the coefficient of COVID-19 was formulated as a bivariate function of post-admission time (i.e., days after admission or length of hospital stay) and calendar time. Similarly as before, the 12 panels of Figure 4.7 present the dynamics of the COVID-19 effect (in log hazard ratio) on three discharge destinations from different perspectives. Panels in the first two rows indicate that patients admitted with COVID-19 were less likely to be discharged to home or to another facility, and more likely to die in hospital or be discharged to hospice than those without COVID-19, especially in the initial phase of the pandemic and over the first 20 days of hospitalization. The COVID-19 effects remained significant with calendar time (the first row), but shrank as the length of stay increased (the second row). Evidence shown in the contour and surface plots (last two rows of Figure 4.7) is consistent with what one would anticipate in the early stage of the pandemic: compared with those admitted without COVID-19, dialysis patients admitted with COVID-19 were associated with a significantly higher risk of early in-hospital death or discharge to hospice, and a significantly lower risk of early discharge to home or another facility. The COVID-19 effects then became less significant until mid-November 2020. After mid-November, the risk of in-hospital death or early discharge to hospice mildly increased among COVID-19 hospitalizations, while the risk of early discharge to another facility decreased substantially among COVID-19 hospitalizations, suggesting a worsening situation toward the end of 2020.

For all three discharge destinations, we performed tests of univariate and bivariate variation of the COVID-19 effects, similarly as in Section 4.6.1. The re-

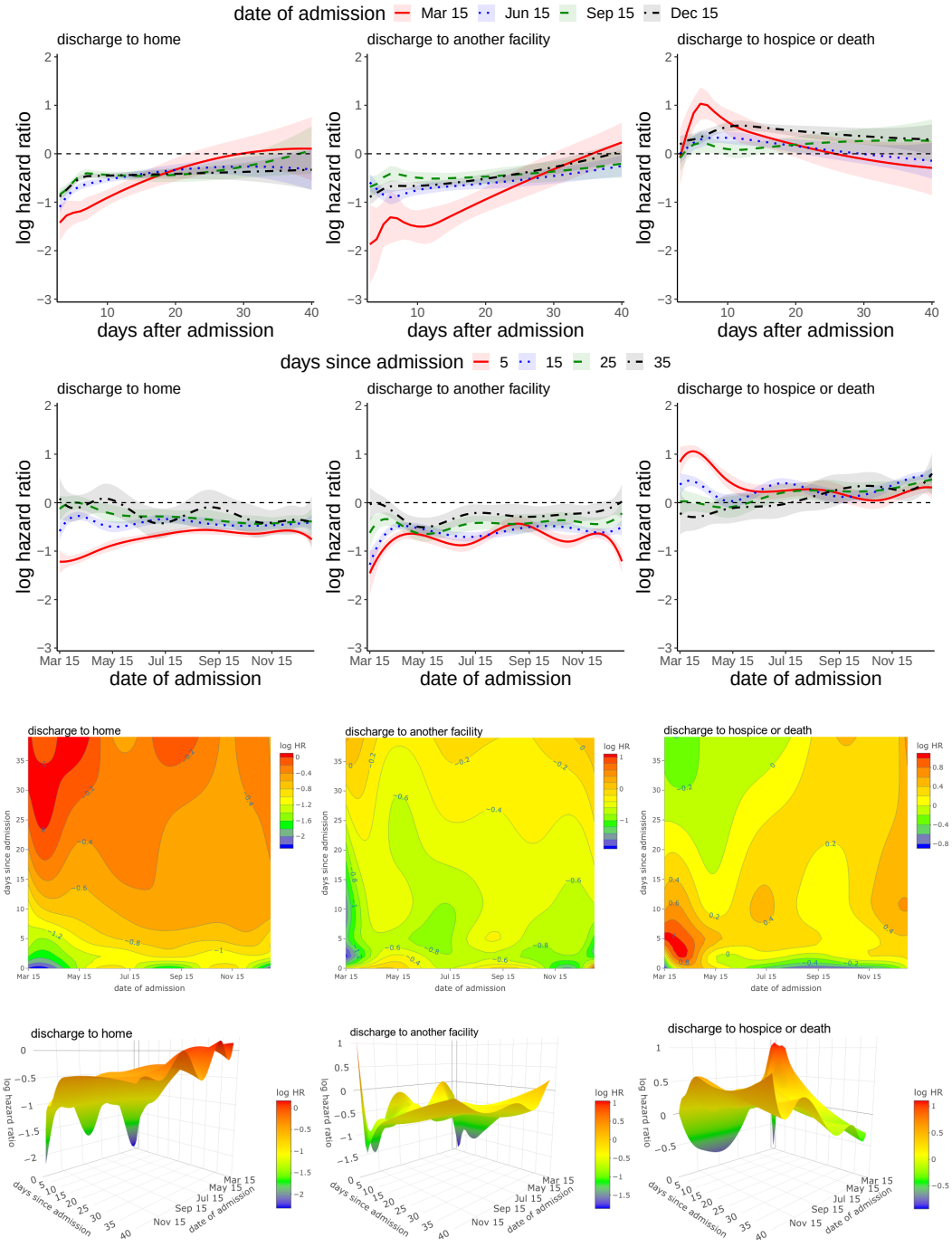


Figure 4.7: Bivariate variation of log hazard ratios with respect to in-hospital COVID-19 diagnosis for discharge status (home, another facility, and hospice/death). Included in the sample were 544,677 unplanned acute-care hospitalizations of 250,940 Medicare beneficiaries on dialysis associated with 2,929 Medicare-certified dialysis facilities in 2020. Ribbons in the top six panels indicate 95% confidence intervals. Panels in the third and fourth rows are contour and surface plots, respectively.

sulting p-values were all less than 0.001, implying that the COVID-19 effects were significantly varying jointly or separately with post-admission and calendar time.

4.7 Discussion

Motivated by our recent investigations into the dynamic impact of COVID-19 on dialysis patients, we have proposed a bivariate varying coefficient model for large-scale competing risks data. This novel model successfully characterizes the variation of COVID-19 effects on both event and calendar timescales. To address the computational challenge arising from fitting the model to the massive data in our applications, we developed an efficient tensor-product proximal Newton algorithm. Further, we introduced difference-based anisotropic penalization to alleviate model overfitting and the unsmoothness of the estimated effect surface. Various methods of cross-validation were considered for parameter tuning purposes. Statistical testing procedures with and without penalization were also designed to examine whether the COVID-10 effect variation was significant across event and calendar time, either jointly or separately. The proposed methods have been comprehensively evaluated through simulation studies and applications to dialysis patients amidst the COVID-19 pandemic.

Although inspired by COVID-19 studies on dialysis patients, the bivariate varying coefficient model can be harnessed in a variety of applications. For instance, among patients with breast cancer, evidence suggests that the racial and ethnic disparities in their cause-specific survival change significantly with post-diagnosis time [136]. The proposed model can be leveraged to examine whether those disparities also change with age at diagnosis, thereby promoting health equity through more customized treatment options.

Multivariate varying coefficient models, as a flexible and granular analytical approach, have been studied in the presence of functional responses [146, 94] or longitudinal outcomes [85, 125]. However, none of these studies has allowed the coefficients to depend on event time in a survival manner, which imposes a higher order of computational complexity than modeling event-time-independent varying coefficients in a multi-dimensional context. In contrast, our proposed model features bivariate effect dependence with both event time and an arbitrary risk factor; the accompanying inference, penalization, and model selection methods also advance the current literature of varying coefficient modeling.

CHAPTER V

Summary and Future Work

This dissertation has introduced three approaches to analyzing massive and complex competing risks data arising from administrative claims and disease registries. In Chapter II, we have developed a discrete time competing risk model for profiling Medicare-certified kidney dialysis facilities based on 30-day unplanned hospital readmissions. Distinct from existing logistic regression models, the new model accounts for event times, and the resulting standardized quality measure is not systematically affected by the rate of competing risks. Next in Chapter III, we have proposed a proximal Newton algorithm that improves the computational efficiency and estimation accuracy of time-varying coefficient modeling for large-scale competing risks data through the introduction of proximal algorithms in convex optimization. Lastly in Chapter IV, a bivariate varying coefficient model has been developed to characterize the multidimensional dynamics of the COVID-19 effect on dialysis patients, accompanied by difference-based anisotropic penalization, tests of variation, and cross-validated model selection methods.

Although the endeavors were motivated by applications in kidney dialysis, cancer survival and COVID-19, the novel competing risk methods can be harnessed in a variety of settings. For instance, the discrete time competing risk model can

be used to examine any type of terminal outcome, as long as the occurrence of the outcome is significantly affected by other types of terminal outcome, e.g., 30-day post-discharge emergency department visit in the presence of post-discharge death. As a second example, one may leverage the bivariate varying coefficient model to study racial and ethnic inequalities in the progression of chronic kidney disease (CKD) to end-stage renal disease, where the disparities may vary with calendar time and the time since the diagnosis of CKD.

In the near future, we intend to extend the research pipeline along several paths. First, evidence suggests that the impact of COVID-19 on the post-discharge prognosis of dialysis patients also varies geographically (Figure 5.1). A competing risk model with spatiotemporally varying coefficients would be an ideal statistical tool for analyzing the variation of the COVID-19 effect across space and time. The literature has seen a growing number of Bayesian [41, 4, 83, 39] and frequentist approaches [145, 113, 82, 81, 65, 64] to modeling spatially varying coefficients for continuous outcomes, with a few studies further considering the spatiotemporal variation [41, 108]. However, there is a paucity of methodological effort devoted to the development of spatially or spatiotemporally varying coefficient models for time-to-event data. Following a frequentist perspective, we propose to bridge this gap by means of the bivariate splines defined on triangulations and their tensor product with univariate splines [67, 107]. Compared with kernel or tensor product smoothing, the proposed method will be free from the "leakage" problem when the spatial data are distributed over irregularly shaped domains with complex boundaries, strong concavities, or interior holes [103]. Given the analytical complexity of our approach, we anticipate the computational challenge to be considerable.

Second, varying coefficient modeling can also be useful in the presence of re-

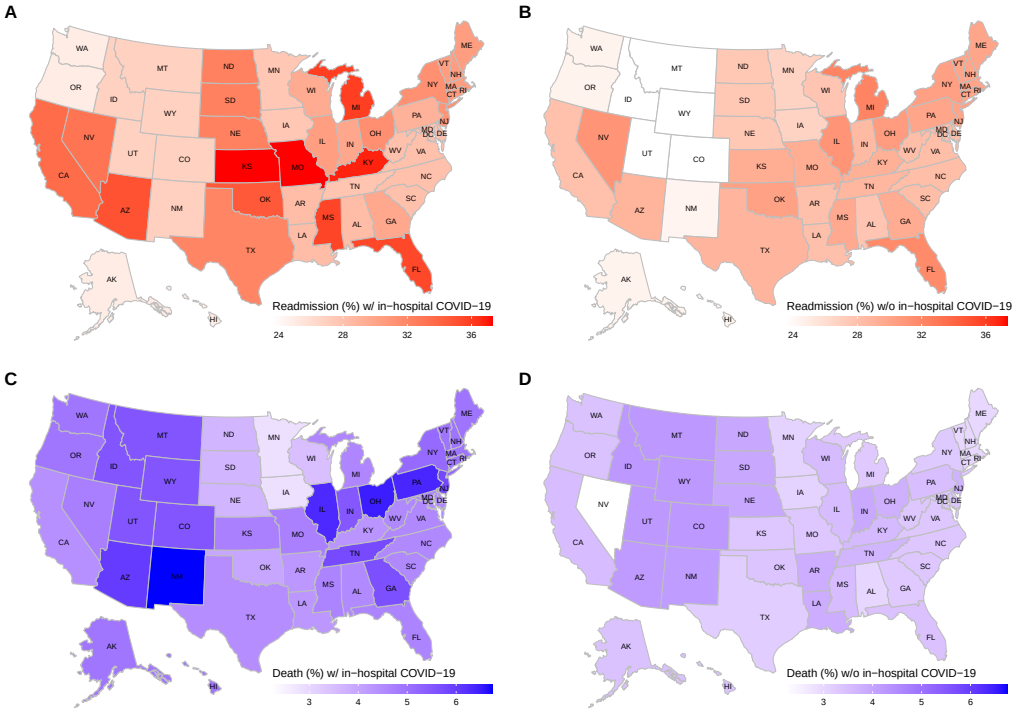


Figure 5.1: Geographic variation of the unplanned readmission and death rates among discharges with and without in-hospital COVID-19 from January 1 to October 31 of 2020. To enhance accuracy, states with limited COVID-19 discharges were combined within each of the nine US Census Bureau-designated divisions.

current events (e.g., hospitalizations) and/or terminal events (e.g., deaths), possibly with staggered subject entry (left truncation). Thus far, there have been only a few articles focused on the temporal variation of coefficients for recurrent events, with or without terminal events. Zhao et al. (2011) [143] considered jointly modeling recurrent and terminal events with time-varying coefficients, which were simply assumed to be piecewise constant. Yu et al. (2014) [142] added a shared random effect (frailty) to the joint event modeling framework with time-varying coefficients. Liu and Guo (2020) [77] pursued a Bayesian framework allowing multitype recurrent events. None of these analyses accounted for the possibility of left truncation, and their applications were restricted to small data sets with about 100 subjects. In our case of hospitalizations among dialysis patients with end-stage renal disease, while some patients were at risk for another hospital admission at a certain time of the year, others were still ineligible for kidney dialysis and should not be counted as at-risk patients. Another overwhelming challenge is that the sample often consists of millions of data records, which renders any general-purpose software implementations infeasible. To better meet the analytical needs, we plan on extending our methodology for competing risks to recurrent and terminal events with left truncation. Since a model with left truncation leads to risk sets non-monotonically varying with time, the load balancing in parallel computing is anticipated to be more challenging than a model with right censoring only. In this context, a solution to optimized workload allocation may involve multi-way number partitioning [105]. Spatial or spatiotemporal variation of coefficients may also be incorporated into the analytical framework.

Lastly, in some applications of time-varying coefficient modeling, it is of particular interest to identify the subset of the domain on which the coefficient exactly equals zero, i.e., a zero-effect (null) region. Despite the growing interest, this topic

remains underexplored. Zhou et al. (2013) [144] developed a shrinkage method which simultaneously detected the null region and estimated the coefficient on the non-null region for functional linear regression models; Yang (2020) [141] proposed a novel soft-thresholded varying coefficient model allowing for null region identification. Both articles were exclusively focused on continuous outcomes. We therefore aim at developing estimation and inference methods of null region detection for competing risks and other time-to-event outcomes.

I embarked on my predoctoral training in biostatistics and scientific computing with a quantitative background in statistics and economics. As my clinical knowledge built up while working at the Kidney Epidemiology and Cost Center, I have become increasingly interested in health services and outcomes research, and aspire to a career that harnesses data-driven quantitative methods to promote population health. The series of competing risk approaches presented here reflects my initial attempt toward the scientific vision. Moving forward, I would like to continue my pursuit along the paths set forth above in the hope of making a difference in public health, one person at a time.

APPENDICES

APPENDIX A

Supplementary Material for Chapter II

A.1 Score and Information of Log-Partial Pseudo-Likelihood L

Let $\mathbf{h}$ denote a vector consisting of

$$h_{ijk} \coloneqq h(\eta_k + \gamma_i + \mathbf{Z}_{ij}^\top(k)\boldsymbol{\beta}),$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, $k = 1, \ldots, t_{ij}$, and let $\dot{h}_{ijk}$ and $\ddot{h}_{ijk}$ denote the first and second order derivatives, respectively. With $\boldsymbol{\theta} \coloneqq (\boldsymbol{\gamma}^\top, \boldsymbol{\eta}^\top, \boldsymbol{\beta}^\top)^\top$, the score and Fisher information of the log-partial pseudo-likelihood are given by

$$\mathcal{U}(\boldsymbol{\theta}) \coloneqq \left(\mathcal{U}^\top(\boldsymbol{\gamma}), \mathcal{U}^\top(\boldsymbol{\eta}), \mathcal{U}^\top(\boldsymbol{\beta})\right)^\top,$$

$$\mathcal{I}(\boldsymbol{\theta}) \coloneqq \begin{pmatrix} \mathcal{I}(\boldsymbol{\gamma}) & \mathcal{I}^\top(\boldsymbol{\eta}, \boldsymbol{\gamma}) & \mathcal{I}^\top(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \mathcal{I}(\boldsymbol{\eta}, \boldsymbol{\gamma}) & \mathcal{I}(\boldsymbol{\eta}) & \mathcal{I}^\top(\boldsymbol{\beta}, \boldsymbol{\eta}) \\ \mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}) & \mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\eta}) & \mathcal{I}(\boldsymbol{\beta}) \end{pmatrix},$$

in which

$$\mathcal{U}(\boldsymbol{\gamma}) = \sum_{k=1}^{\tau} \left(\mathcal{U}_{1\cdot k}, \ldots, \mathcal{U}_{m\cdot k}\right)^\top, \mathcal{I}(\boldsymbol{\gamma}) = \sum_{k=1}^{\tau} \operatorname{diag}(\mathcal{I}_{1\cdot k}, \ldots, \mathcal{I}_{m\cdot k}),$$

$$\mathcal{U}(\boldsymbol{\eta}) = \sum_{i=1}^{m} \left(\mathcal{U}_{i\cdot 1}, \ldots, \mathcal{U}_{i\cdot \tau}\right)^\top, \mathcal{I}(\boldsymbol{\eta}) = \sum_{i=1}^{m} \operatorname{diag}(\mathcal{I}_{i\cdot 1}, \ldots, \mathcal{I}_{i\cdot \tau}),$$

$$\mathcal{U}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{\tau} \mathcal{U}_{ijk} \mathbf{Z}_{ij}(k), \mathcal{I}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{\tau} \mathcal{I}_{ijk} \mathbf{Z}_{ij}(k) \mathbf{Z}_{ij}^\top(k),$$

$$\mathcal{I}(\boldsymbol{\eta},\boldsymbol{\gamma}) = \begin{pmatrix} \mathcal{I}_{1\cdot 1} & \cdots & \mathcal{I}_{m\cdot 1} \\ \vdots & \ddots & \vdots \\ \mathcal{I}_{1\cdot \tau} & \cdots & \mathcal{I}_{m\cdot \tau} \end{pmatrix},$$

$$\mathcal{I}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{k=1}^{\tau} \left\{ \sum_{j=1}^{n_1} \mathcal{I}_{1jk}\mathbf{Z}_{1j}(k), \ldots, \sum_{j=1}^{n_m} \mathcal{I}_{mjk}\mathbf{Z}_{mj}(k) \right\},$$

$$\mathcal{I}(\boldsymbol{\beta},\boldsymbol{\eta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\mathcal{I}_{ij1}\mathbf{Z}_{ij}(1), \ldots, \mathcal{I}_{ij\tau}\mathbf{Z}_{ij}(\tau) \right),$$

with

$$\mathcal{U}_{ijk} \coloneqq \frac{\dot{h}_{ijk}\{\Delta N_{ij}^1(k) - Y_{ij}(k)h_{ijk}\}}{h_{ijk}(1-h_{ijk})}, \quad \mathcal{I}_{ijk} \coloneqq \frac{Y_{ij}(k)\dot{h}_{ijk}^2}{h_{ijk}(1-h_{ijk})}.$$

Throughout the expressions above, a subscripted dot denotes summation over that subscript. For convenience, let $\mathcal{D}(\boldsymbol{\gamma}) \coloneqq [\mathcal{I}_{1\cdot\cdot}^{-1}, \ldots, \mathcal{I}_{m\cdot\cdot}^{-1}]^\top$, the vector of diagonal elements of $\mathcal{I}^{-1}(\boldsymbol{\gamma})$, and

$$\mathcal{I}_{11} \coloneqq \mathcal{I}(\boldsymbol{\gamma}),\ \mathcal{I}_{21} \coloneqq \begin{pmatrix} \mathcal{I}(\boldsymbol{\eta},\boldsymbol{\gamma}) \\ \mathcal{I}(\boldsymbol{\beta},\boldsymbol{\gamma}) \end{pmatrix},\ \mathcal{I}_{12} = \mathcal{I}_{21}^\top,\ \mathcal{I}_{22} \coloneqq \begin{pmatrix} \mathcal{I}(\boldsymbol{\eta}) & \mathcal{I}^\top(\boldsymbol{\beta},\boldsymbol{\eta}) \\ \mathcal{I}(\boldsymbol{\beta},\boldsymbol{\eta}) & \mathcal{I}(\boldsymbol{\beta}) \end{pmatrix}.$$

A.2 Blockwise Inversion Newton Algorithm: Technical Details

Let $\circ$ denote the entrywise product, and let $l \in \{0\} \cup \mathbb{N}$ index iterations. The Blockwise Inversion Newton algorithm is sketched as Algorithm 4. Computing the Newton step $\Delta\boldsymbol{\theta} = \mathcal{I}^{-1}(\boldsymbol{\theta})\mathcal{U}(\boldsymbol{\theta})$ in Lines 8 and 9 is a dominant bottleneck. By the blockwise inversion formula [9], we have

$$\mathcal{I}^{-1} = \begin{pmatrix} \mathcal{I}_{11}^{-1} + \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\mathcal{S}^{-1}\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & -\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\mathcal{S}^{-1} \\ -\mathcal{S}^{-1}\mathcal{I}_{21}\mathcal{I}_{11}^{-1} & \mathcal{S}^{-1} \end{pmatrix}, \tag{A.1}$$

where $\mathcal{S} \coloneqq \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}$ is the Schur complement of $\mathcal{I}_{11} = \mathcal{I}(\boldsymbol{\gamma})$. Observing the four blocks of $\mathcal{I}^{-1}(\boldsymbol{\theta})$, we record $\mathcal{J}_1$, $\mathcal{S}^{-1}$ and $\mathcal{J}_2$ (Lines 5–7, Algorithm 4) at each iteration to avoid redundant computing. Specifically, $\mathcal{J}_1$ defined as a matrix product,

can instead be computed by multiplying each column of $\mathcal{I}_{21}$ with the corresponding element of $\mathcal{D}(\boldsymbol{\gamma})$, a vector of the diagonal elements of $\mathcal{I}^{-1}(\boldsymbol{\gamma})$. This trick brings down the time complexity from $O(m^2(\tau+r))$ to $O(m(\tau+r))$. In total, inverting $\mathcal{I}(\boldsymbol{\theta})$ via (A.1) costs $O(m(\tau+r)^2+(\tau+r)^3)$, much less than $O((m+\tau+r)^3)$ using a naive Newton–Raphson algorithm given that $m \gg \tau + r$.

Algorithm 4: Blockwise Inversion Newton (BIN)

initialize $l \leftarrow 0$ and $\boldsymbol{\theta}^{(0)} = \mathbf{0}$
set $s \in (0, 0.5)$, $t \in (0.5, 1)$ and $\epsilon > 0$
do
  $d \leftarrow 1$
  $\mathcal{J}_1^{(l)} = \mathcal{I}_{21}^{(l)} \left\{\mathcal{I}_{11}^{(l)}\right\}^{-1}$
  $\mathcal{S}^{(l)} = \mathcal{I}_{22}^{(l)} - \mathcal{J}_1^{(l)} \left\{\mathcal{I}_{21}^{(l)}\right\}^{\top}$
  $\mathcal{J}_2^{(l)} = \left\{\mathcal{S}^{(l)}\right\}^{-1} \mathcal{J}_1^{(l)}$
  $\Delta\boldsymbol{\gamma}^{(l)} = \mathcal{D}(\boldsymbol{\gamma}^{(l)}) \circ \mathcal{U}(\boldsymbol{\gamma}^{(l)}) + \left\{\mathcal{J}_2^{(l)}\right\}^{\top} \left\{\mathcal{J}_1^{(l)} \mathcal{U}(\boldsymbol{\gamma}^{(l)}) - \begin{pmatrix} \mathcal{U}(\boldsymbol{\eta}^{(l)}) \\ \mathcal{U}(\boldsymbol{\beta}^{(l)}) \end{pmatrix}\right\}$
  $\begin{pmatrix} \Delta\boldsymbol{\eta}^{(l)} \\ \Delta\boldsymbol{\beta}^{(l)} \end{pmatrix} = \left\{\mathcal{S}^{(l)}\right\}^{-1} \begin{pmatrix} \mathcal{U}(\boldsymbol{\eta}^{(l)}) \\ \mathcal{U}(\boldsymbol{\beta}^{(l)}) \end{pmatrix} - \mathcal{J}_2^{(l)} \mathcal{U}(\boldsymbol{\gamma}^{(l)})$
  while $\ell(\boldsymbol{\theta}^{(l)} + d\Delta\boldsymbol{\theta}^{(l)}) < \ell(\boldsymbol{\theta}^{(l)}) + sd\mathcal{U}^{\top}(\boldsymbol{\theta}^{(l)})\Delta\boldsymbol{\theta}^{(l)}$ **do**
    $d \leftarrow td$
  end while
while $\|\mathbf{h}^{(l)} - \mathbf{h}^{(l-1)}\|_{\infty} \geq \epsilon$

A.3 Justifying $\hat{\boldsymbol{\theta}}$ as a GEE Estimator

We redefine discharge j as discharge l of patient p to introduce patient-level indexing. Let $n^{(i)}$ denote the patient count of facility i, let n_{ip} denote the discharge count of patient p in that facility, and let $\bar{n}$ be the maximum of all n_{ip}'s. To ease notation, for $i = 1, \ldots, m$, $p = 1, \ldots, n^{(i)}$, $l = 1, \ldots, n_{ip}$, and $k = 1, \ldots, \tau$, we define

$$\mathbf{U}_{ipl}(k) \coloneqq (u_{iplk}(1), \ldots, u_{iplk}(m))^{\top},$$

$$\mathbf{W}_{ipl}(k) \coloneqq (w_{iplk}(1), \ldots, w_{iplk}(\tau))^{\top},$$

where $u_{iplk}(q) = 1(q = j)$ and $w_{iplk}(q) = 1(q = k)$. Then we have

$$h_{iplk} = h(\eta_k + \gamma_i + \mathbf{Z}_{ipl}^{\top}(k)\boldsymbol{\beta}) = h(\mathbf{X}_{ipl}^{\top}(k)\boldsymbol{\theta}),$$

in which $\mathbf{X}_{ipl}(k) \coloneqq (\mathbf{U}_{ipl}^\top(k), \mathbf{W}_{ipl}^\top(k), \mathbf{Z}_{ipl}^\top(k))^\top$. Further, we let

$$\begin{aligned}
\mathbf{D}_{ip} &= (\dot{h}_{ip11}\mathbf{X}_{ip1}(1), \ldots, \dot{h}_{ip1\tau}\mathbf{X}_{ip1}(\tau), \ldots, \dot{h}_{ipn_{ip}1}\mathbf{X}_{ipn_{ip}}(1), \ldots, \dot{h}_{ipn_{ip}\tau}\mathbf{X}_{ipn_{ip}}(\tau))^\top, \\
\Delta\mathbf{N}_{ip}^1 &\coloneqq (\Delta N_{ip1}^1(1), \ldots, \Delta N_{ip1}^1(\tau), \ldots, \Delta N_{ipn_{ip}}^1(1), \ldots, \Delta N_{ipn_{ip}}^1(\tau))^\top, \\
\mathbf{Y}_{ip} &\coloneqq \mathrm{diag}(Y_{ip1}(1), \ldots, Y_{ip1}(\tau), \ldots, Y_{ipn_{ip}}(1), \ldots, Y_{ipn_{ip}}(\tau)), \\
\mathbf{h}_{ip} &\coloneqq (h_{ip11}, \ldots, h_{ip1\tau}, \ldots, h_{ipn_{ip}1}, \ldots, h_{ipn_{ip}\tau})^\top, \\
\mathbf{V}_{ip} &\coloneqq \mathrm{diag}(v(h_{ip11}), \ldots, v(h_{ip1\tau}), \ldots, v(h_{ipn_{ip}1}), \ldots, v(h_{ipn_{ip}\tau})), \\
\mathbf{b}_{ip} &\coloneqq (b(h_{ip11}), \ldots, b(h_{ip1\tau}), \ldots, b(h_{ipn_{ip}1}), \ldots, b(h_{ipn_{ip}\tau}))^\top,
\end{aligned}$$

where h_{iplk} is a shorthand of $h(\eta_k + \gamma_i + \mathbf{Z}_{ij}^\top(k)\boldsymbol{\beta})$, and b and v are known functions given by

$$b(h) \coloneqq \sqrt{h(1-h)}, \quad v(h) \coloneqq \frac{\dot{h}^2}{h(1-h)},$$

with $\dot{h}$ representing the first-order derivative of h.

Following the GEE framework, we have a system of unbiased estimating equations

$$\sum_{i=1}^{m}\sum_{p=1}^{n^{(i)}} \mathbf{D}_{ip}^\top \mathbf{V}_{ip}^{-1}(\Delta\mathbf{N}_{ip}^1 - \mathbf{Y}_{ip}\mathbf{h}_{ip}) = \mathbf{0}, \tag{A.2}$$

in which an independent working correlation matrix has been assumed. With some algebra, system (A.2) reduces to

$$\phi^{-1}\sum_{i=1}^{m}\sum_{p=1}^{n^{(i)}}\sum_{l=1}^{n_{ip}}\sum_{k=1}^{\tau} \mathcal{U}_{iplk}\mathbf{X}_{ipl}(k) = \mathbf{0},$$

which is easily seen to be equivalent to $\mathcal{U}(\boldsymbol{\theta}) = \mathbf{0}$. This implies that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is a solution to (A.2), i.e., $\hat{\boldsymbol{\theta}}$ is a GEE estimator under the working independence assumption.

A.4 Assumptions on the Stabilized Robust Variance Estimator

As in a marginal model, we make three assumptions on the stabilized robust variance estimator.

Assumption A.1. The conditional mean of $\Delta N^1_{ipl}(k)$ (i.e., cause-specific hazard of readmission) depends on the covariates $\mathbf{Z}_{ipl}(k)$ via a known function, that is,

$$\mathrm{E}\{\Delta N^1_{ipl}(k)|T_{ipl} \geq k, \mathbf{Z}_{ipl}(k)\} = \lambda_{ipl}(k) = h_{iplk}. \tag{A.3}$$

Assumption A.2. The conditional variance of $\Delta N^1_{ipl}(k)$ depends on the conditional mean h_{iplk} in (A.3) according to

$$\mathrm{Var}\{\Delta N^1_{ipl}(k)|T_{ipl} \geq k, \mathbf{Z}_{ipl}(k)\} = v(h_{iplk})\phi,$$

where $\phi > 0$ is an unknown scale parameter.

Assumption A.3. For any $n \in \{1, \ldots, \bar{n}\}$, all patients who have a τnth-order leading principal submatrix of the correlation matrix of $\Delta\mathbf{N}^1_{ip}$ share a common structure for that submatrix.

Compared with Pan (2001) [90], Assumption A.3 extends the notion of common correlation structure to patient-level clustering with unequal discharge counts. An underlying justification is that for those patients who have experienced at least n discharges, their levels of exposure to additional UHRs should be similar, regardless of the different discharge counts they would eventually have. This assumption allows inter-patient discharge pooling to be independent of patient-specific discharge count, which, implicitly, is a random variable for every patient.

A.5 Stabilized Robust Score Test: Technical Details

For patient p in facility i, let $\mathbf{h}_{ip}$ be a vector of UHR hazards $h_{iplk} \coloneqq h(\eta_k + \gamma_i + \mathbf{Z}^\top_{ipl}(k)\boldsymbol{\beta})$, let $\mathbf{Y}_{ip}$ denote a diagonal matrix of at-risk indicators $Y_{ipl}(k)$,

let $\mathbf{V}_{ip}$ be a diagonal matrix with variance functions $v(h_{iplk})$ on the diagonal, where $v(h) \coloneqq \dot{h}^2/[h(1-h)]$ with $\dot{h}$ being the first derivative of h, and let $\mathbf{b}_{ip}$ be an entrywise transformation of $\mathbf{h}_{ip}$ by $b(h) \coloneqq \sqrt{h(1-h)}$. Furthermore, we let $n^{(i)}$ denote the patient count of facility i, let n_{ip} denote the discharge count of patient p in that facility, and let $\bar{n}$ be the maximum of all n_{ip}'s. As typically assumed, we adopt a working independence covariance structure of $\Delta\mathbf{N}^1_{ip}$, the vector of UHR outcomes of patient p in facility i. Under this condition, $\hat{\boldsymbol{\theta}}$ is a generalized estimating equation (GEE) estimator. Detailed derivation is available in Web Appendix B of the Supporting Information. With the notation in Section A.3, the integrated correlation matrix $\mathbf{M}$ is defined as

$$\mathbf{M} \coloneqq \boldsymbol{\Pi} \circ \mathbf{A} \coloneqq \boldsymbol{\Pi} \circ \sum_{i=1}^{m} \sum_{p=1}^{n^{(i)}} \mathbf{A}_{ip}, \tag{A.4}$$

where matrix $\boldsymbol{\Pi} \coloneqq (\pi_{rc}) \in \mathbb{R}^{\tau\bar{n}\times\tau\bar{n}}$ with $\pi_{rc} \coloneqq |\{(i,p) : n_{ip} \geq \lceil \max(r,c)/\tau \rceil\}|^{-1}$, and $\mathbf{A}_{ip} \in \mathbb{R}^{\tau\bar{n}\times\tau\bar{n}}$ is zero everywhere except at its τn_{ip}th-order leading principal submatrix (LPS) $\mathbf{A}^{(n_{ip})}_{ip} \coloneqq \mathbf{V}^{-1/2}_{ip}(\Delta\mathbf{N}^1_{ip} - \mathbf{Y}_{ip}\mathbf{h}_{ip})(\Delta\mathbf{N}^1_{ip} - \mathbf{Y}_{ip}\mathbf{h}_{ip})^\top\mathbf{V}^{-1/2}_{ip}$. The matrix $\mathbf{M}$ can be thought of as an aggregation of patient-specific correlation information $\mathbf{A}^{(n_{ip})}_{ip}$ of unequal dimensions, then downscaled by the counts of patients who contribute to the specific entries of $\mathbf{A}$. Figure A.1 illustrates how $\boldsymbol{\Pi}$ is applied to $\mathbf{A}$ to obtain $\mathbf{M}$. With the matrix $\mathbf{M}$ defined as in (A.4), the stabilized robust estimator of $\mathrm{Var}(\hat{\gamma}_i)$ can be written as

$$\widehat{\Sigma}_{\gamma_i} \coloneqq \widehat{W}^{-1}_{\gamma_i}\widehat{V}_{\gamma_i}\widehat{W}^{-1}_{\gamma_i}, \tag{A.5}$$

where

$$\widehat{V}_{\gamma_i} \coloneqq \hat{\phi}^{-2}\sum_{p=1}^{n^{(i)}} \hat{\mathbf{b}}^\top_{ip}\widehat{\mathbf{M}}^{(n_{ip})}\hat{\mathbf{b}}_{ip}, \quad \widehat{W}_{\gamma_i} \coloneqq \hat{\phi}^{-1}\sum_{p=1}^{n^{(i)}} \hat{\mathbf{b}}^\top_{ip}\mathbf{Y}_{ip}\hat{\mathbf{b}}_{ip},$$

$$\hat{\phi} \coloneqq N^{-1}\sum_{i=1}^{m}\sum_{p=1}^{n^{(i)}}\sum_{l=1}^{n_{ip}}\sum_{k=1}^{\tau} \frac{\{\Delta N^1_{ipl}(k) - Y_{ipl}(k)\hat{h}_{iplk}\}^2}{v(\hat{h}_{iplk})},$$

$N \coloneqq \sum_{i=1}^{m} \sum_{p=1}^{n^{(i)}} \sum_{l=1}^{n_{ip}} \sum_{k=1}^{\tau} Y_{ipl}(k) - m - \tau - p + 1$, $\mathbf{M}^{(n)}$ is the τnth-order LPS of $\mathbf{M}$, and carets indicate evaluations at $\hat{\boldsymbol{\theta}}$.

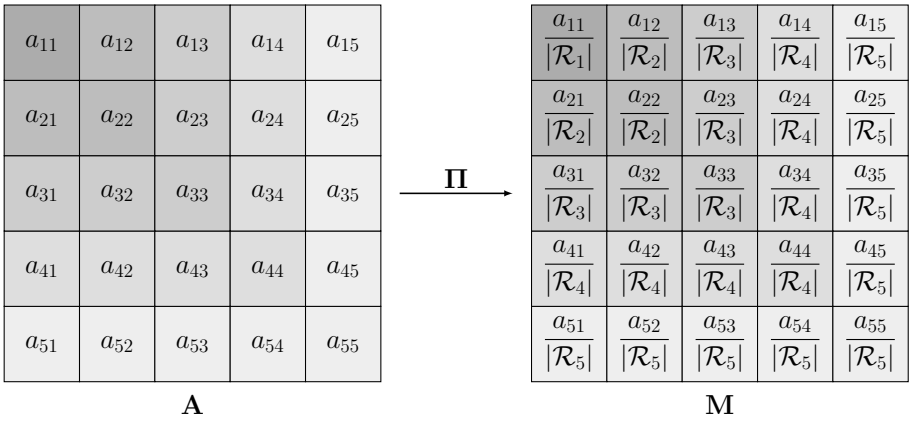


Figure A.1: Illustration of constructing $\mathbf{M}$ as an entrywise product of $\mathbf{A}$ and the scale matrix $\boldsymbol{\Pi}$. For simplicity, we consider $\tau = 1$ and $\bar{n} = 5$. A lighter tint of gray indicates a smaller set $\mathcal{R}_n \coloneqq \{(i,p) : n_{ip} \geq n\}$ of patients at risk for at least n discharges, $n = 1, \ldots, \bar{n}$.

A.6 Alternative Tests

A generalized Wald test statistic is given by

$$T_i^{\mathrm{RW}} \coloneqq \frac{\hat{\gamma}_i - \hat{\gamma}_{\mathrm{M}}}{\sqrt{\widehat{\Sigma}_{\gamma_i}}},$$

which has the same asymptotic distribution as T_i^{RS}.

If the competing risk model is correctly specified, a model-based score test statistic under the null hypothesis $H_{0i} : \gamma_i = \gamma_{\mathrm{M}}$ can be expressed as

$$T_i^{\mathrm{MS}} \frac{\mathcal{U}_i(\tilde{\boldsymbol{\gamma}}_i)}{\sqrt{\mathcal{I}_i(\tilde{\boldsymbol{\gamma}}_i)}},$$

where $\mathcal{I}_i(\tilde{\boldsymbol{\gamma}}_i)$ is the ith diagonal element of the information matrix $\mathcal{I}(\tilde{\boldsymbol{\gamma}}_i)$ evaluated at $\tilde{\boldsymbol{\theta}}_i$. Likewise, a model-based Wald test statistic is given by

$$T_i^{\mathrm{MW}} \coloneqq \sqrt{\mathcal{I}_i(\hat{\boldsymbol{\gamma}})}(\hat{\gamma}_i - \hat{\gamma}_{\mathrm{M}}),$$

where $\mathcal{I}_i(\hat{\boldsymbol{\gamma}})$ is the ith diagonal element of $\mathcal{I}(\hat{\boldsymbol{\gamma}})$ evaluated at $\hat{\boldsymbol{\theta}}$. Under certain regularity conditions, T_i^{MS} and T_i^{MW} also have an asymptotic standard normal distribution.

A.7 Simulation Details

As acknowledged in the literature [95, 90, 29], with a logit link and clustered binary responses, it is unlikely to build a model in which subjects with the same number of repeated measurements share a common correlation matrix structure. Since our inference framework bears a resemblance to a marginal model, the previous observation carries over to our simulation setting. Nonetheless, Assumptions A.1 to A.3 in Section A.4 can be approximately satisfied by assuming the following misspecified model with random effect:

$$\mathrm{E}\{\Delta N^1_{ipl}(k) | T_{ipl} \geq k, \mathbf{Z}_{ipl}(k), \varepsilon_{ipl}\} = h(\eta_k + \gamma_i + \mathbf{Z}^\top_{ipl}(k)\boldsymbol{\beta} + \varepsilon_{ipl}),$$

where $\boldsymbol{\varepsilon}_{ip} = [\varepsilon_{ip1}, \ldots, \varepsilon_{ipn_{ip}}]^\top \sim \mathcal{N}_{n_{ip}}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ being an exchangeable covariance matrix with marginal variance σ^2 and correlation ρ. When h is a standard logistic function (corresponding to the logit link), we have

$$\mathrm{E}\{\Delta N^1_{ipl}(k) | T_{ipl} \geq k, \mathbf{Z}_{ipl}(k)\} \approx \psi_\sigma(\eta_k + \gamma_i + \mathbf{Z}^\top_{ipl}(k)\boldsymbol{\beta}),$$

where

$$\psi_\sigma(u) \coloneqq \kappa \cdot \Phi\left(\frac{u}{\sqrt{\xi_1 + \sigma^2}}\right) + (1 - \kappa) \cdot \Phi\left(\frac{u}{\sqrt{\xi_2 + \sigma^2}}\right),$$

with $\kappa = 0.4353$, $\xi_1 = 2.2967^2$, and $\xi_2 = 1.3017^2$, and the last so-called two-probit approximation is due to Demidenko (2013) [28]. Figure A.2 displays absolute errors of two-probit approximation with respect to the standard logistic function under different values of σ. In particular, $\sigma = 0.3$ strikes a balance between introducing correlation and approximating the logistic function, with a maximum absolute error of 0.0043.

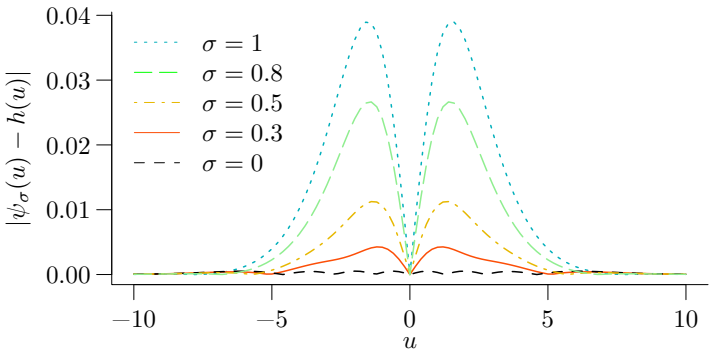


Figure A.2: Absolute errors of two-probit approximation relative to standard logistic function under different values of σ.

A.8 Application Details

The readmissions data are derived from an extensive national ESRD patient database garnered from multiple data systems operated by federal public health agencies in the United States. Significant sources include the Renal Management Information System, CROWNWeb facility-reported clinical and administrative data (extracted from CMS-2728 Medical Evidence Form, CMS-2746 Death Notification Form, and CMS-2744 Annual Facility Survey Form), Medicare Enrollment Database, Medicare claims data from Standard Analytic Files, transplant data from the Scientific Registry of Transplant Recipients, nursing home Minimum Data Set, provider survey and certification data from Quality Improvement and Evaluation System Business Intelligence Center, and the Dialysis Facility Compare.

Applying exclusion criteria (such as patients with a primary diagnosis of cancer, mental health, or rehabilitation, discharged against medical advice, or hospitalized at Prospective Payment System-exempt cancer hospitals) to qualifying discharges taking place between January 1 and December 31, 2018, there were 541,769 discharges (257,860 patients) from 6,937 Medicare-certified dialysis facilities included in analysis, each facility with at least 11 discharges. The 6,937 facilities had a wide

variety of discharge counts from 11 (52 facilities) to 842 (1 facility) with a mean of 78.10, readmission counts from 0 (19) to 264 (1) with a mean of 20.58, and competing event counts from 0 (707) to 76 (1) with a mean of 3.72. Among them, 10% had at most 24 discharges or 5 readmissions, and 26.35% had at most 1 competing risk. Patient-specific discharge counts spanned from 1 (143,704 patients) to 12 (139 patients), with 75% of patients having at most 2 discharges. The observed 30-day facility-specific readmission rates ranged from 0% (19 facilities) to 73.33% (1 facility), with an overall readmission rate of 26.35%; the observed 30-day facility-specific rates of competing risks varied from 0% (707) to 23.08% (1), with an overall rate of 4.76%. We consider 74 predictors, including age, sex, body mass index, years on dialysis, status of Medicare Advantage Plans at discharge, length (days) of index hospitalization, diabetic status, past-year nursing home status at discharge, past-year prevalent comorbidities and high-risk conditions at discharge (using Agency for Healthcare Research and Quality Clinical Classifications Software ICD-10 diagnosis categories). Days (4 to 30) to the first event after discharge are available as times to events. If a patient switches to another facility prior to a readmission, then the readmission is attributed to the dialysis facility at the time of discharge.

A.9 Supplementary Figures

Figure A.3 implies that different link functions give rise to similar distribution of SRRs in our application setting. Figure A.4 demonstrates that using estimates from the constrained model refitting procedures yield similar score test statistics as those without model refitting, regardless of which variance estimator is considered. Figure A.5 shows the difference in SRR versus average at-risk time, stratified by facility-specific discharge counts.

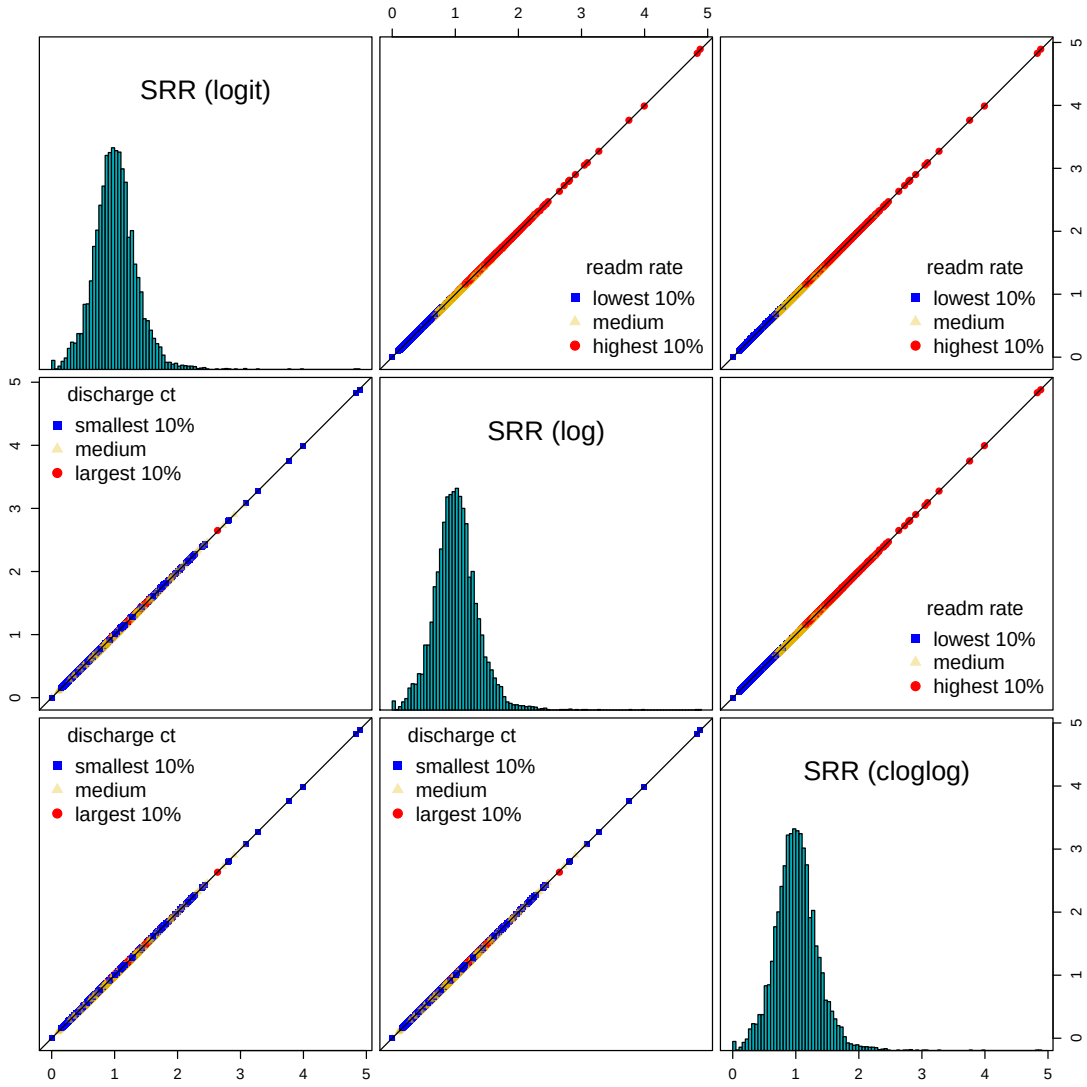


Figure A.3: SRR from competing risk models with different link functions. Histograms are in the diagonal panels. Facilities are stratified by readmission rate or discharge count. Dashed lines represent 2.5% and 97.5% quantiles of the standard normal distribution. 45-degree lines in solid black.

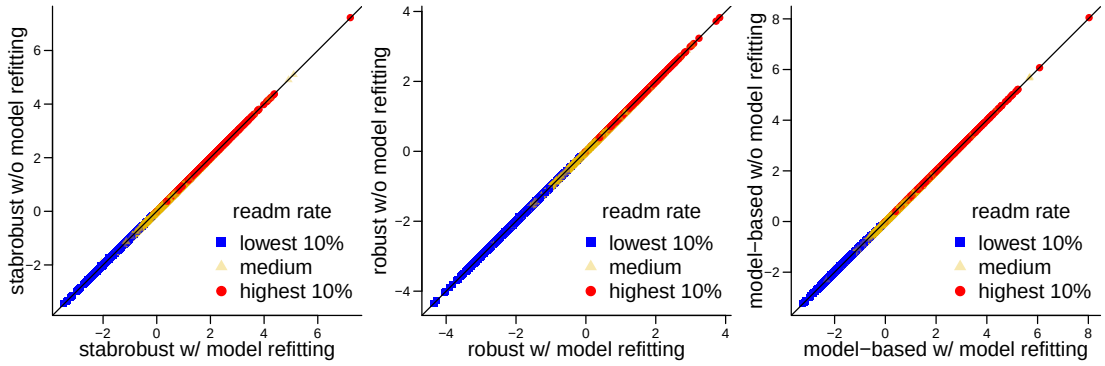


Figure A.4: Score test statistics with versus without constrained model refitting using different variance estimators. "stabrobust", "robust" and "model" correspond to test statistics with the stabilized robust, classical robust and model-based variance estimators, respectively. Facilities are stratified by readmission rate. 45-degree lines in solid black.

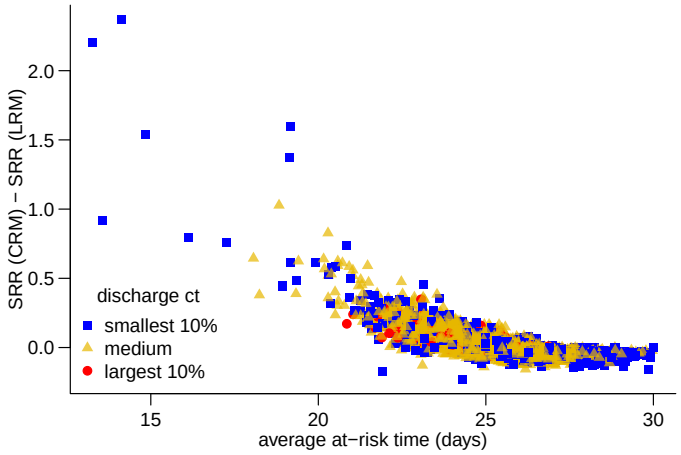


Figure A.5: Difference in SRR versus average at-risk time, stratified by facility-specific discharge counts. SRRs under the CRM and LRM are computed based on expression (8) in the article and He et al. (2013) [49], respectively. The at-risk time of a discharge is defined as the earlier of the time to the first event and the end of follow-up (30 days).

APPENDIX B

Supplementary Material for Chapter III

B.1 Gradient and Information of Log-Partial Likelihood (3.4)

To derive the gradient $\nabla \ell_j(\boldsymbol{\gamma}_j)$ and Hessian matrix $\nabla^2 \ell_j(\boldsymbol{\gamma}_j)$ of $\ell_j(\boldsymbol{\gamma}_j)$, we define

$$S_{ij}^{(u)}(\boldsymbol{\gamma}_j, X_i) \coloneqq \sum_{r \in R(X_i)} \exp\{[\mathbf{Z}_r \otimes \mathbf{B}(X_i)]^\top \boldsymbol{\gamma}_j\} \mathbf{Z}_r^{\odot u}, \quad u = 0, 1, 2,$$

where for a vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^{\odot 0} \coloneqq 1$, $\mathbf{v}^{\odot 1} \coloneqq \mathbf{v}$, and $\mathbf{v}^{\odot 2} \coloneqq \mathbf{v}\mathbf{v}^\top$. The gradient $\nabla \ell_j(\boldsymbol{\gamma}_j)$ and Hessian $\nabla^2 \ell_j(\boldsymbol{\gamma}_j)$ of $\ell_j(\boldsymbol{\gamma}_j)$ are hence given by

$$\nabla \ell_j(\boldsymbol{\gamma}_j) = \frac{1}{n} \sum_{i=1}^{n} \Delta_{ij} \left\{ \mathbf{Z}_i - \overline{\mathbf{Z}}_{ij}(\boldsymbol{\gamma}_j, X_i) \right\} \otimes \mathbf{B}(X_i), \tag{B.1}$$

$$\nabla^2 \ell_j(\boldsymbol{\gamma}_j) = -\frac{1}{n} \sum_{i=1}^{n} \Delta_{ij} \mathbf{V}_{ij}(\boldsymbol{\gamma}_j, X_i) \otimes \left\{ \mathbf{B}(X_i) \mathbf{B}^\top(X_i) \right\}, \tag{B.2}$$

in which

$$\overline{\mathbf{Z}}_{ij}(\boldsymbol{\gamma}_j, X_i) \coloneqq \frac{S_{ij}^{(1)}(\boldsymbol{\gamma}_j, X_i)}{S_{ij}^{(0)}(\boldsymbol{\gamma}_j, X_i)}, \quad \mathbf{V}_{ij}(\boldsymbol{\gamma}_j, X_i) \coloneqq \frac{S_{ij}^{(2)}(\boldsymbol{\gamma}_j, X_i)}{S_{ij}^{(0)}(\boldsymbol{\gamma}_j, X_i)} - \overline{\mathbf{Z}}_{ij}^{\odot 2}(\boldsymbol{\gamma}_j, X_i).$$

B.2 Proofs of Lemmas, Propositions and Theorems

Proposition B.1. *Let* $\mathbf{prox}_{\lambda \ell}$ *be a proximal operator of* $\lambda \ell$ *as in (4) with the maximand* $g(\boldsymbol{\gamma}) \coloneqq \ell(\boldsymbol{\gamma}) - \frac{1}{2\lambda} \|\boldsymbol{\gamma} - \mathbf{v}\|_2^2$. *Then* $\exists\, m > 0$ *such that* $h(\boldsymbol{\gamma}) \coloneqq g(\boldsymbol{\gamma}) + \frac{m}{2} \|\boldsymbol{\gamma}\|_2^2$ *is concave and the maximizer of* g *is unique.*

Proof. Given any $\mu \in [0,1]$ and $\boldsymbol{\gamma}, \boldsymbol{\alpha} \in \mathrm{dom}(g)$, observe that when $m \leq 1/\lambda$,

$$\begin{aligned}\mu h(\boldsymbol{\gamma}) + (1-\mu)h(\boldsymbol{\alpha}) - h(\mu\boldsymbol{\gamma} + (1-\mu)\boldsymbol{\alpha}) &= \mu\ell(\boldsymbol{\gamma}) + (1-\mu)\ell(\boldsymbol{\alpha}) - \ell(\mu\boldsymbol{\gamma} + (1-\mu)\boldsymbol{\alpha}) \\ &\quad + \left(\frac{m}{2} - \frac{1}{2\lambda}\right)\left[\mu\|\boldsymbol{\gamma}\|_2^2 + (1-\mu)\|\boldsymbol{\alpha}\|_2^2 - \|\mu\boldsymbol{\gamma} + (1-\mu)\boldsymbol{\alpha}\|_2^2\right] \leq 0,\end{aligned}$$

by the concavity of ℓ and convexity of the Euclidean norm $\|\cdot\|_2$. Thus g is strongly and hence strictly concave, which further implies that g has a unique maximizer. ■

Lemma B.2. *Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then a direction $\boldsymbol{\mu} \in \mathbb{R}^d$ satisfies $\nabla\ell(\boldsymbol{\gamma})^\top\boldsymbol{\mu} > 0$ at $\boldsymbol{\gamma}$ if and only if there exists a symmetric and positive definite matrix $\mathbf{M}$ such that $\boldsymbol{\mu} = \mathbf{M}^{-1}\nabla\ell(\boldsymbol{\gamma})$.*

Proof. Suppose $\boldsymbol{\mu} = \mathbf{M}^{-1}\nabla\ell(\boldsymbol{\gamma})$ where $\mathbf{M}$ is positive definite. Then $\mathbf{M}^{-1}$ is positive definite as well since eigenvalues of $\mathbf{M}^{-1}$ are reciprocals of eigenvalues of $\mathbf{M}$ by spectral decomposition. Thus $\nabla\ell(\boldsymbol{\gamma})^\top\boldsymbol{\mu} = \nabla\ell(\boldsymbol{\gamma})^\top\mathbf{M}^{-1}\nabla\ell(\boldsymbol{\gamma}) > 0$ whenever $\nabla\ell(\boldsymbol{\gamma}) \neq \mathbf{0}$. Conversely, suppose $\mathbf{d}^\top\boldsymbol{\mu} > 0$ where $\mathbf{d} \coloneqq \nabla\ell(\boldsymbol{\gamma})$. We claim that

$$\mathbf{M} \coloneqq \mathbf{I} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}} + \frac{\mathbf{d}\mathbf{d}^\top}{\mathbf{d}^\top\boldsymbol{\mu}}$$

is positive definite. Observe that $\mathbf{I} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}}$ is symmetric and idempotent, and $\frac{\mathbf{d}\mathbf{d}^\top}{\mathbf{d}^\top\boldsymbol{\mu}}$ is symmetric and positive semidefinite. Then $\mathbf{M}$ is positive semidefinite. Pick an arbitrary $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ with

$$\mathbf{x}^\top\left(\mathbf{I} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}}\right)\mathbf{x} = 0 \Leftrightarrow \left(\mathbf{I} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}}\right)\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{x} \in \mathrm{null}\left(\mathbf{I} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}}\right) = \mathrm{range}\left(\frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}}\right).$$

That $\mathbf{x} \in \mathrm{range}\left(\frac{\boldsymbol{\mu}\boldsymbol{\mu}^\top}{\boldsymbol{\mu}^\top\boldsymbol{\mu}}\right)$ implies that $\exists\, \mathbf{y} \in \mathbb{R}^d$ such that $\mathbf{x} = a\boldsymbol{\mu}$ with $a \coloneqq \frac{\boldsymbol{\mu}^\top\mathbf{y}}{\boldsymbol{\mu}^\top\boldsymbol{\mu}} \neq 0$. Then

$$\mathbf{x}^\top\left(\frac{\mathbf{d}\mathbf{d}^\top}{\mathbf{d}^\top\boldsymbol{\mu}}\right)\mathbf{x} = \frac{a^2\boldsymbol{\mu}^\top\mathbf{d}\mathbf{d}^\top\boldsymbol{\mu}}{\mathbf{d}^\top\boldsymbol{\mu}} = a^2\mathbf{d}^\top\boldsymbol{\mu} > 0$$

implies that $\mathbf{x}^\top\mathbf{M}\mathbf{x} > 0$ for any $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, that is, $\mathbf{M}$ is positive definite. Furthermore, we have $\mathbf{M}\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu} + \mathbf{d} = \mathbf{d} = \nabla\ell(\boldsymbol{\gamma})$, i.e., $\boldsymbol{\mu} = \mathbf{M}^{-1}\nabla\ell(\boldsymbol{\gamma})$. ■

Proposition B.3. *Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and bounded above on $\mathbb{R}^d$. Assume that $\exists\, \boldsymbol{\gamma}^{(0)} \in \mathbb{R}^d$ such that the superlevel set $G \coloneqq \{\boldsymbol{\gamma} \in \mathbb{R}^d : \ell(\boldsymbol{\gamma}) \geq \ell(\boldsymbol{\gamma}^{(0)})\}$ is convex and compact with a nonempty interior $\operatorname{int}(G)$. If the sequence $\{\boldsymbol{\gamma}^{(s)}\}_{s=1}^{\infty}$ is defined recursively by $\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} + \nu \Delta\boldsymbol{\gamma}^{(s)}$, where $\Delta\boldsymbol{\gamma}^{(s)}$ satisfies $\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)} > 0$ at $\boldsymbol{\gamma}^{(s)}$, and $\nu > 0$ satisfies*

$$\ell(\boldsymbol{\gamma}^{(s)} + \nu\Delta\boldsymbol{\gamma}^{(s)}) \geq \ell(\boldsymbol{\gamma}^{(s)}) + \phi\nu\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}, \tag{B.3}$$

$$\ell(\boldsymbol{\gamma}^{(s)} + \nu\Delta\boldsymbol{\gamma}^{(s)}) \leq \ell(\boldsymbol{\gamma}^{(s)}) + \psi\nu\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}, \tag{B.4}$$

with $\phi \in (0,1)$ and $\psi \in (\phi, 1)$, then $\{\ell(\boldsymbol{\gamma}^{(s)})\}_{s=0}^{\infty}$ converges and

$$\lim_{s\to\infty} \frac{\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2} = 0. \tag{B.5}$$

Proof. By the Heine–Cantor theorem, the continuity of $\nabla\ell$ on a compact set G implies uniform continuity of $\nabla\ell$. For fixed $\boldsymbol{\gamma} \in G$ and $\boldsymbol{\mu}$ satisfying $\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} > 0$, define

$$R(\boldsymbol{\gamma}, \boldsymbol{\mu}) \coloneqq \{\nu > 0 : \phi\nu\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} \leq \ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) - \ell(\boldsymbol{\gamma}) \leq \psi\nu\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu}\}.$$

Trivially, $\forall\, \nu \in R(\boldsymbol{\gamma}, \boldsymbol{\mu})$, $\boldsymbol{\gamma} + \nu\boldsymbol{\mu} \in G$, which implies that $\{\boldsymbol{\gamma}^{(s)}\}_{s=1}^{\infty} \subset G$. Now we claim that $R(\boldsymbol{\gamma}, \boldsymbol{\mu})$ contains a nontrivial interval. Since ℓ is bounded above by M, and $\phi\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} > 0$, $\exists\, \nu_u > 0$ such that $\forall\, \nu \geq \nu_u$, $\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) \leq M < \ell(\boldsymbol{\gamma}) + \phi\nu\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu}$. By the definition of directional derivative, $\exists\, \epsilon > 0$ such that $\forall\, \nu \in (0, \epsilon \wedge \nu_u)$,

$$\frac{\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) - \ell(\boldsymbol{\gamma})}{\nu} - \nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} > -(1-\phi)\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} \Leftrightarrow \ell(\boldsymbol{\gamma}+\nu\boldsymbol{\mu}) > \ell(\boldsymbol{\gamma})+\phi\nu\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu}.$$

Since $\nu \mapsto \ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) - \phi\nu\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu}$ is a continuous mapping, by the intermediate value theorem, $\exists\, \nu_1 \in (0, \nu_u)$ such that $\ell(\boldsymbol{\gamma} + \nu_1\boldsymbol{\mu}) = \ell(\boldsymbol{\gamma}) + \phi\nu_1\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu}$. With the same argument, $\exists\, \nu_2 > 0$ such that $\ell(\boldsymbol{\gamma} + \nu_2\boldsymbol{\mu}) = \ell(\boldsymbol{\gamma}) + \psi\nu_2\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu}$. Once again, the intermediate value theorem implies that $\exists\, \tau > 0$ such that

$$\phi\tau\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu} < \ell(\boldsymbol{\gamma} + \tau\boldsymbol{\mu}) - \ell(\boldsymbol{\gamma}) < \psi\tau\nabla\ell^\top(\boldsymbol{\gamma})\boldsymbol{\mu},$$

which by the continuity of $\nu \mapsto [\ell(\boldsymbol{\gamma} + \nu\boldsymbol{\mu}) - \ell(\boldsymbol{\gamma})]/\nu$ implies that there exists a nonempty open neighborhood of τ contained in $R(\boldsymbol{\gamma}, \boldsymbol{\mu})$.

Obviously, $\{\ell(\boldsymbol{\gamma}^{(s)})\}_{s=1}^{\infty}$ is an increasing sequence bounded above and hence converges. To show (B.5) by contradiction, suppose $\exists\, \epsilon > 0$ and a subsequence $\{\boldsymbol{\gamma}^{(s_r)}\}_{r=1}^{\infty}$ such that

$$\frac{\nabla \ell^{\top}(\boldsymbol{\gamma}^{(s_r)})\Delta\boldsymbol{\gamma}^{(s_r)}}{\|\Delta\boldsymbol{\gamma}^{(s_r)}\|_2} \geq \epsilon, \quad \forall\, r \in \mathbb{N}. \tag{B.6}$$

Condition (B.3) implies that

$$\begin{aligned}\ell(\boldsymbol{\gamma}^{(s_{r+1})}) - \ell(\boldsymbol{\gamma}^{(s_r)}) &\geq \ell(\boldsymbol{\gamma}^{(s_r)} + \nu^{(s_r)}\Delta\boldsymbol{\gamma}^{(s_r)}) - \ell(\boldsymbol{\gamma}^{(s_r)}) \\ &\geq \phi\nu^{(s_r)}\nabla\ell^{\top}(\boldsymbol{\gamma}^{(s_r)})\Delta\boldsymbol{\gamma}^{(s_r)} \geq \phi\nu^{(s_r)}\|\Delta\boldsymbol{\gamma}^{(s_r)}\|_2\epsilon, \quad \forall\, r \in \mathbb{N},\end{aligned}$$

which further implies that

$$\lim_{r\to\infty} \nu^{(s_r)}\|\Delta\boldsymbol{\gamma}^{(s_r)}\|_2 = 0 \tag{B.7}$$

given that $\{\ell(\boldsymbol{\gamma}^{(s_r)})\}_{r=1}^{\infty}$ converges.

In addition, Condition (B.4) implies that $\forall\, r \in \mathbb{N}$,

$$(1-\psi)\nu^{(s_r)}\nabla\ell^{\top}(\boldsymbol{\gamma}^{(s_r)})\Delta\boldsymbol{\gamma}^{(s_r)} \leq \nu^{(s_r)}\nabla\ell^{\top}(\boldsymbol{\gamma}^{(s_r)})\Delta\boldsymbol{\gamma}^{(s_r)} + \ell(\boldsymbol{\gamma}^{(s_r)}) - \ell(\boldsymbol{\gamma}^{(s_r)} + \nu^{(s_r)}\Delta\boldsymbol{\gamma}^{(s_r)}).$$

By Taylor's theorem and the Cauchy–Schwarz inequality, $\exists\, \xi \in (0,1)$ such that

$$\begin{aligned}&(1-\psi)\nu^{(s_r)}\nabla\ell^{\top}(\boldsymbol{\gamma}^{(s_r)})\Delta\boldsymbol{\gamma}^{(s_r)} \\ &\qquad \leq \nu^{(s_r)}\nabla\ell^{\top}(\boldsymbol{\gamma}^{(s_r)})\Delta\boldsymbol{\gamma}^{(s_r)} + \ell(\boldsymbol{\gamma}^{(s_r)}) - \ell(\boldsymbol{\gamma}^{(s_r)} + \nu^{(s_r)}\Delta\boldsymbol{\gamma}^{(s_r)}) \\ &\qquad = \nu^{(s_r)}\left[\nabla\ell(\boldsymbol{\gamma}^{(s_r)}) - \nabla\ell(\boldsymbol{\gamma}^{(s_r)} + \xi\nu^{(s_r)}\Delta\boldsymbol{\gamma}^{(s_r)})\right]^{\top}\Delta\boldsymbol{\gamma}^{(s_r)} \\ &\qquad\qquad \leq \|\nabla\ell(\boldsymbol{\gamma}^{(s_r)}) - \nabla\ell(\boldsymbol{\gamma}^{(s_r)} + \xi\nu^{(s_r)}\Delta\boldsymbol{\gamma}^{(s_r)})\|_2\nu^{(s_r)}\|\Delta\boldsymbol{\gamma}^{(s_r)}\|_2.\end{aligned}$$

By (B.6), (B.7), and the uniform continuity of $\nabla\ell$, $\exists\, \xi \in (0,1)$ such that as $r \to \infty$,

$$0 < (1-\psi)\epsilon \leq \|\nabla\ell(\boldsymbol{\gamma}^{(s_r)}) - \nabla\ell(\boldsymbol{\gamma}^{(s_r)} + \xi\nu^{(s_r)}\Delta\boldsymbol{\gamma}^{(s_r)})\|_2 \to 0,$$

a contradiction. Thus (B.5) holds true. ■

Theorem B.4. *Let ℓ_j assume the log-partial likelihood (4) in Section 2 with an initial iterate $\boldsymbol{\gamma}_j^{(0)}$, and let $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ be a sequence of iterates defined by Line 20 of Algorithm 1, where $\Delta\boldsymbol{\gamma}_j^{(s)}$ is given by Line 16, and $\nu > 0$ is determined by the Armijo–Goldstein conditions (7) and (8) in Section 3.2 with $\phi \in (0, 0.5)$ and $\psi \in (0.5, 1)$. If Assumptions 1 and 2 hold, then $\{\ell_j(\boldsymbol{\gamma}_j^{(s)})\}_{s=0}^{\infty}$ converges and $\lim_{s\to\infty} \|\nabla\ell_j(\boldsymbol{\gamma}_j^{(s)})\|_2 = 0$.*

Proof. By assumption, Condition (B.5) holds by Proposition B.3. $\forall\, s \in \mathbb{N}$, let $\mathbf{M}_{(s)} \coloneqq \mathbf{I}/\lambda_s - \nabla^2\ell_j(\boldsymbol{\gamma}_j^{(s)})$, symmetric and positive definite. Spectral decomposition implies that $\|\mathbf{M}_{(s)}^{1/2}\|_2 = \|\mathbf{M}_{(s)}\|_2^{1/2}$ and $\|\mathbf{M}_{(s)}^{-1/2}\|_2 = \|\mathbf{M}_{(s)}\|_2^{-1/2}$. It follows from Assumption 2 that

$$\begin{aligned}
\frac{\nabla\ell_j^{\top}(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}}{\|\nabla\ell(\boldsymbol{\gamma}_j^{(s)})\|_2\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2} &= \frac{(\Delta\boldsymbol{\gamma}_j^{(s)})^{\top}\mathbf{M}_{(s)}\Delta\boldsymbol{\gamma}_j^{(s)}}{\|\mathbf{M}_{(s)}\Delta\boldsymbol{\gamma}_j^{(s)}\|_2\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2} \\
&\geq \frac{(\Delta\boldsymbol{\gamma}_j^{(s)})^{\top}\mathbf{M}_{(s)}^{1/2}\mathbf{M}_{(s)}^{1/2}\Delta\boldsymbol{\gamma}_j^{(s)}}{\|\mathbf{M}_{(s)}\|_2\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2^2} \\
&= \frac{\|\mathbf{M}_{(s)}^{1/2}\Delta\boldsymbol{\gamma}_j^{(s)}\|_2^2}{\|\mathbf{M}_{(s)}\|_2\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2^2} \\
&\geq \frac{\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2^2}{\|\mathbf{M}_{(s)}^{-1/2}\|_2^2\|\mathbf{M}_{(s)}\|_2\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2^2} \\
&= \frac{1}{\|\mathbf{M}_{(s)}^{-1}\|_2\|\mathbf{M}_{(s)}\|_2} \geq \frac{1}{\kappa},
\end{aligned}$$

that is,

$$\frac{\nabla\ell^{\top}(\boldsymbol{\gamma}_j^{(s)})\Delta\boldsymbol{\gamma}_j^{(s)}}{\|\Delta\boldsymbol{\gamma}_j^{(s)}\|_2} \geq \frac{\|\nabla\ell(\boldsymbol{\gamma}_j^{(s)})\|_2}{\kappa}, \quad \forall\, s \in \mathbb{N},$$

which by (B.5) implies $\lim_{s\to\infty} \|\nabla\ell_j(\boldsymbol{\gamma}_j^{(s)})\|_2 = 0$. ■

Lemma B.5. *If $\ell : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then $\forall\, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$,*

$$\|\nabla\ell(\mathbf{y}) - \nabla\ell(\mathbf{z}) - \nabla^2\ell(\mathbf{x})(\mathbf{y} - \mathbf{z})\|_2 \leq \sup_{0\leq\xi\leq 1} \|\nabla^2\ell(\mathbf{z} + \xi(\mathbf{y} - \mathbf{z})) - \nabla^2\ell(\mathbf{x})\|_2\|\mathbf{y} - \mathbf{z}\|_2.$$

See, for example, Ortega and Rheinboldt (1970) [89] for a proof.

Proposition B.6. *Let $\ell : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Assume that $\exists\, \boldsymbol{\gamma}^{(0)} \in \mathbb{R}^d$ such that the superlevel set $G \coloneqq \{\boldsymbol{\gamma} \in \mathbb{R}^d : \ell(\boldsymbol{\gamma}) \geq \ell(\boldsymbol{\gamma}^{(0)})\}$ is convex and compact with a nonempty interior $\operatorname{int}(G)$, and that the sequence $\{\boldsymbol{\gamma}^{(s)}\}_{s=1}^{\infty}$ defined recursively by $\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} + \nu \Delta\boldsymbol{\gamma}^{(s)}$ converges to $\boldsymbol{\gamma}^*$, where $\nabla^2 \ell(\boldsymbol{\gamma}^*)$ is negative definite, $\Delta\boldsymbol{\gamma}^{(s)}$ satisfies $\nabla \ell^\top(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)} > 0$ at $\boldsymbol{\gamma}^{(s)}$, and $\nu > 0$ satisfies* (B.3) *and* (B.4) *with $\phi \in (0, 0.5)$ and $\psi \in (0.5, 1)$. If*

$$\lim_{s \to \infty} \frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2} = 0, \tag{B.8}$$

then

(i) $\exists\, s_0 \in \mathbb{N}$ such that $\forall\, s \geq s_0$, $\nu = 1$ satisfies (B.3) *and* (B.4)*;*

(ii) $\nabla \ell(\boldsymbol{\gamma}^*) = \mathbf{0}$; and

(iii) $\{\boldsymbol{\gamma}^{(s)}\}_{s=0}^{\infty}$ converges superlinearly to $\boldsymbol{\gamma}^*$ provided that $\forall\, s \geq s_0$, $\nu = 1$ for some $s_0 \in \mathbb{N}$.

Proof. The Cauchy–Schwarz inequality implies that

$$\begin{aligned}
\frac{\nabla \ell^\top(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2} &= \frac{(\Delta\boldsymbol{\gamma}^{(s)})^\top [\nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}] - (\Delta\boldsymbol{\gamma}^{(s)})^\top \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2} \\
&\geq \frac{-\|\Delta\boldsymbol{\gamma}^{(s)}\|_2 \|\nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2} - \frac{(\Delta\boldsymbol{\gamma}^{(s)})^\top \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2} \\
&\geq -\frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2} - \lambda_{\max}(\nabla^2 \ell(\boldsymbol{\gamma}^{(s)})),
\end{aligned}$$

where $\lambda_{\max}(\mathbf{M})$ denotes the greatest eigenvalue of a matrix $\mathbf{M}$. By the assumption on ℓ, the map $\boldsymbol{\gamma} \mapsto \lambda_{\max}(\nabla^2 \ell(\boldsymbol{\gamma}))$ is continuous. By (B.8) and the negative definiteness of $\nabla^2 \ell(\boldsymbol{\gamma}^*)$, for sufficiently large $s_1 \in \mathbb{N}$, $\exists\, \xi > 0$ such that

$$\frac{\nabla \ell^\top(\boldsymbol{\gamma}^{(s)}) \Delta\boldsymbol{\gamma}^{(s)}}{\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2} \geq \xi, \quad \forall\, s \geq s_1. \tag{B.9}$$

It follows from (B.5) that

$$\lim_{s \to \infty} \|\Delta\boldsymbol{\gamma}^{(s)}\|_2 = 0. \tag{B.10}$$

By Taylor's theorem, $\exists\, \bar{\boldsymbol{\gamma}} \coloneqq \boldsymbol{\gamma}^{(s)} + \xi \Delta\boldsymbol{\gamma}^{(s)}$ with $\xi \in (0,1)$ such that

$$\begin{aligned}
&\ell(\boldsymbol{\gamma}^{(s)} + \Delta\boldsymbol{\gamma}^{(s)}) - \ell(\boldsymbol{\gamma}^{(s)}) - \frac{1}{2}\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)} \\
&= \frac{1}{2}[\nabla\ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2\ell(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}]^\top\Delta\boldsymbol{\gamma}^{(s)} + \frac{1}{2}(\Delta\boldsymbol{\gamma}^{(s)})^\top[\nabla^2\ell(\bar{\boldsymbol{\gamma}}) - \nabla^2\ell(\boldsymbol{\gamma}^{(s)})]\Delta\boldsymbol{\gamma}^{(s)} \\
&\geq -\left[\frac{\|\nabla\ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2\ell(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}\|_2}{2\|\Delta\boldsymbol{\gamma}^{(s)}\|_2} + \frac{1}{2}\|\nabla^2\ell(\bar{\boldsymbol{\gamma}}) - \nabla^2\ell(\boldsymbol{\gamma}^{(s)})\|_2\right]\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2.
\end{aligned}$$

It follows from (B.8), (B.10), and the continuity of $\boldsymbol{\gamma} \mapsto \nabla^2\ell(\boldsymbol{\gamma})$ that for sufficiently large $s_2 \in \mathbb{N}$,

$$\begin{aligned}
&\ell(\boldsymbol{\gamma}^{(s)} + \Delta\boldsymbol{\gamma}^{(s)}) - \ell(\boldsymbol{\gamma}^{(s)}) - \frac{1}{2}\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)} \\
&\qquad\qquad \geq -(1/2 - \phi)\xi\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2 \geq (\phi - 1/2)\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}, \quad \forall\, s \geq s_2,
\end{aligned}$$

the second inequality due to (B.9). Thus $\forall\, s \geq s_2$, $\nu = 1$ satisfies (B.3). Likewise, since $\psi > 1/2$, for sufficiently large $s_3 \in \mathbb{N}$,

$$\begin{aligned}
&\ell(\boldsymbol{\gamma}^{(s)} + \Delta\boldsymbol{\gamma}^{(s)}) - \ell(\boldsymbol{\gamma}^{(s)}) - \frac{1}{2}\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)} \\
&\qquad\qquad \leq (\psi - 1/2)\xi\|\Delta\boldsymbol{\gamma}^{(s)}\|_2^2 \leq (\psi - 1/2)\nabla\ell^\top(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}, \quad \forall\, s \geq s_3,
\end{aligned}$$

which satisfies (B.4) and hence Part (i) holds true.

To show Part (ii), note that (B.10) implies

$$\lim_{s\to\infty} \|\nabla\ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2\ell(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}\|_2 = 0.$$

By the triangle inequality and continuity of $\boldsymbol{\gamma} \mapsto \nabla^2\ell(\boldsymbol{\gamma})$, as $s \to \infty$,

$$\|\nabla\ell(\boldsymbol{\gamma}^{(s)})\|_2 \leq \|\nabla\ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2\ell(\boldsymbol{\gamma}^{(s)})\Delta\boldsymbol{\gamma}^{(s)}\|_2 + \|\nabla^2\ell(\boldsymbol{\gamma}^{(s)})\|_2\|\Delta\boldsymbol{\gamma}^{(s)}\|_2 \to 0,$$

that is, Part (ii) is true.

To show Part (iii), suppose without loss of generality that $\forall\, s \in \mathbb{N}$, $\nu = 1$ satisfies Conditions (B.3) and (B.4). By the triangle inequality, Lemma B.5 and

(B.8), as $s \to \infty$,

$$\begin{aligned}
\frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s+1)})\|_2}{\|\Delta \boldsymbol{\gamma}^{(s)}\|_2} &= \frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s+1)}) - \nabla \ell(\boldsymbol{\gamma}^{(s)}) - \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta \boldsymbol{\gamma}^{(s)} + \nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta \boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta \boldsymbol{\gamma}^{(s)}\|_2} \\
&\leq \frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s+1)}) - \nabla \ell(\boldsymbol{\gamma}^{(s)}) - \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta \boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta \boldsymbol{\gamma}^{(s)}\|_2} \\
&\qquad + \frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta \boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta \boldsymbol{\gamma}^{(s)}\|_2} \\
&\leq \sup_{0 \leq \xi \leq 1} \|\nabla^2 \ell(\boldsymbol{\gamma}^{(s)} + \xi \Delta \boldsymbol{\gamma}^{(s)}) - \nabla^2 \ell(\boldsymbol{\gamma}^{(s)})\|_2 + \frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s)}) + \nabla^2 \ell(\boldsymbol{\gamma}^{(s)}) \Delta \boldsymbol{\gamma}^{(s)}\|_2}{\|\Delta \boldsymbol{\gamma}^{(s)}\|_2} \\
&\to 0,
\end{aligned}$$

given that $\lim_{s \to \infty} \|\xi \Delta \boldsymbol{\gamma}^{(s)}\|_2 = 0$ by (B.10). Likewise, by Part (ii), the triangle inequality, and Lemma B.5, for sufficiently large s_4, $\exists\, \eta > 0$ such that

$$\begin{aligned}
&\|\nabla \ell(\boldsymbol{\gamma}^{(s+1)})\|_2 = \|\nabla \ell(\boldsymbol{\gamma}^{(s+1)}) - \nabla \ell(\boldsymbol{\gamma}^*)\|_2 \\
&\geq \|\nabla^2 \ell(\boldsymbol{\gamma}^*)(\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*)\|_2 - \|\nabla \ell(\boldsymbol{\gamma}^{(s+1)}) - \nabla \ell(\boldsymbol{\gamma}^*) - \nabla^2 \ell(\boldsymbol{\gamma}^*)(\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*)\|_2 \\
&\geq \left[\frac{1}{\|\nabla^2 \ell(\boldsymbol{\gamma}^*)^{-1}\|_2} - \sup_{0 \leq \xi \leq 1} \|\nabla^2 \ell(\boldsymbol{\gamma}^{(s)} + \xi(\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*)) - \nabla^2 \ell(\boldsymbol{\gamma}^*)\|_2 \right] \|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2 \\
&\geq \eta \|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2, \quad \forall\, s \geq s_4.
\end{aligned}$$

By the arguments above and triangle inequality, as $s \to \infty$, we have

$$\frac{\eta \rho^{(s)}}{1 + \rho^{(s)}} = \frac{\eta \|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2}{\|\boldsymbol{\gamma}^{(s)} - \boldsymbol{\gamma}^*\|_2 + \|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2} \leq \frac{\|\nabla \ell(\boldsymbol{\gamma}^{(s+1)})\|_2}{\|\Delta \boldsymbol{\gamma}^{(s)}\|_2} \to 0,$$

where $\rho^{(s)} \coloneqq \|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^*\|_2 / \|\boldsymbol{\gamma}^{(s)} - \boldsymbol{\gamma}^*\|_2$. It follows that $\lim_{s \to \infty} \rho^{(s)} = 0$ and hence by Definition 2 in Section 3.2, Part (iii) holds. ■

Theorem B.7. *Let ℓ_j assume the log-partial likelihood (4) in Section 2 with an initial iterate $\boldsymbol{\gamma}_j^{(0)}$, and let $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ be a sequence of iterates defined by Line 20 of Algorithm 1, where $\Delta \boldsymbol{\gamma}_j^{(s)}$ is given by Line 16, and $\nu > 0$ is determined by the Armijo–Goldstein conditions (7) and (8) in Section 3.2 with $\phi \in (0, 0.5)$ and $\psi \in (0.5, 1)$. In addition, assume that $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=1}^{\infty}$ converges to $\boldsymbol{\gamma}_j^*$ with a negative definite $\nabla^2 \ell_j(\boldsymbol{\gamma}_j^*)$. If Assumptions 1 and 3 hold, then (1) $\exists\, s_0 \in \mathbb{N}$ such that $\forall\, s \geq s_0$,*

$\nu = 1$ satisfies the Armijo–Goldstein conditions; (2) $\nabla \ell_j(\boldsymbol{\gamma}_j^*) = \mathbf{0}$; and (3) $\{\boldsymbol{\gamma}_j^{(s)}\}_{s=0}^{\infty}$ converges superlinearly to $\boldsymbol{\gamma}_j^*$ provided that $\forall\, s \geq s_0$, $\nu = 1$ for some $s_0 \in \mathbb{N}$.

Proof. By the assumptions, we have

$$\begin{aligned}\nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)}) \Delta \boldsymbol{\gamma}_j^{(s)} &= \nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) + \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)}) \left(\mathbf{I}/\lambda_s - \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)}) \right)^{-1} \nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) \\ &= (\mathbf{I}/\lambda_s) \left(\mathbf{I}/\lambda_s - \nabla^2 \ell_j(\boldsymbol{\gamma}_j^{(s)}) \right)^{-1} \nabla \ell_j(\boldsymbol{\gamma}_j^{(s)}) = \Delta \boldsymbol{\gamma}_j^{(s)}/\lambda_s,\end{aligned}$$

which implies that (B.8) holds. Then the conclusions follow from Proposition B.6. ■

B.3 Proximal Newton algorithm versus its parallelization

To illustrate the advantage of the parallelized ProxiN, in Figure B.1, we reported its speedup and efficiency with respect to the serial ProxiN with varying thread counts. The setting was similar to the one in Table 1 of the manuscript, except that the sample size increased from 1,000 to 1,000,000. As more threads were involved, the speedup curves went up and efficiency curves went down. When the sample size was 1,000, the speedup and efficiency were at the lowest level since the overhead of invoking the shared-memory parallelization accounted for the vast majority of the computational cost. However, when the sample size increased to 10,000, the speedup and efficiency reached 12 and 0.75, respectively, with 16 threads being employed. This indicated a massive advantage of the parallelized ProxiN algorithm over the serial version. As the sample size increased further to 1 million, the speedup and efficiency dropped to 7.43 and 0.464, respectively, in the presence of 16 threads. Therefore, we recommend that the parallelized ProxiN algorithm be preferred to the serial ProxiN especially when the sample size is beyond 10,000. The speed improvement is substantial even on a general-purpose computer workstation with a limited number of threads.

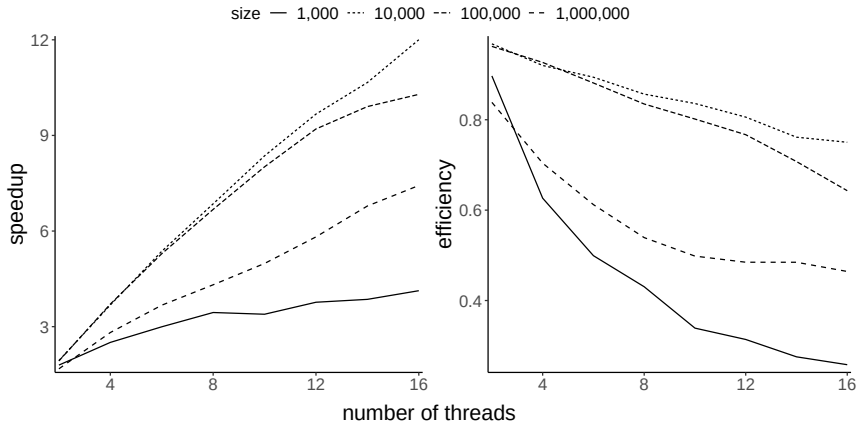


Figure B.1: Speedup and efficiency of the parallelized proximal Newton algorithm. Experiments were conducted using simulated data on an Intel® Xeon® Gold 6254 quad-processor with max frequency 4 GHz and RAM 576 GB. The sample size varied from 1,000 to 1,000,000. The runtime was taken as an average of the duration of 10 runs with a fixed number of threads. The speedup was defined as the ratio of the runtime of the serial proximal Newton algorithm to the runtime of the parallelized version, given a certain number of threads. The efficiency was defined as the speedup divided by the number of threads. See Casanova et al. (2008) [12] for a detailed account.

B.4 Evaluating parallelized ProxiN using breast and prostate cancer data

We evaluated the shared-memory parallelization of ProxiN with a varied number of threads and presented the speedup and efficiency in Figure B.2. As the number of threads increased from 2 to 16, the speedup of the parallelization relative to the serial proximal algorithm grew from 1.56 to 10.79. As more threads were involved in the computation, fewer tasks were assigned to a single thread and the load distribution became less even. Therefore, the per thread proportion of parallelization overhead rose, resulting in an overall efficiency decline from 78.18% to 67.47%. Theoretical upper bounds of speedup and efficiency curves were also depicted.

Using the prostate cancer data, we also assessed the shared-memory parallelization of ProxiN, with results shown in Figure B.3. As the number of threads grew from 2 to 16, the speedup increased from 1.75 to 7.42, while the efficiency declined from 87.51% to 46.41%. Comparing the results with those using the breast cancer

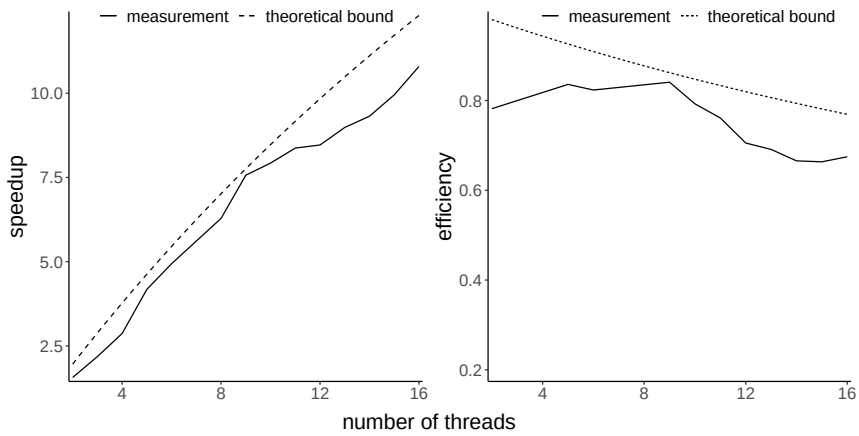


Figure B.2: Speedup and efficiency of the parallelized proximal Newton algorithm. Experiments were conducted using the SEER breast cancer data on an Intel® Xeon® Gold 6254 quad-processor with max frequency 4 GHz and RAM 576 GB. The runtime was taken as an average of the duration of 10 runs with a fixed number of threads. The speedup was defined as the ratio of the runtime of the serial proximal Newton algorithm to the runtime of the parallelized version, given a certain number of threads. The efficiency was defined as the speedup divided by the number of threads. See Casanova et al. (2008) [12] for a detailed account. Theoretical bounds of speedup and efficiency were also obtained according to Amdahl's law [1]. Let $\delta \in (0, 1)$ be the fraction of serial runtime of parallelizable code and c the number of threads used for parallelization. Then the speedup is bounded by $c/(\delta + c - c\delta)$, and the efficiency by $1/(\delta + c - c\delta)$. In this case, the fraction δ was around 98%. For reference, the serial proximal Newton algorithm took 159.10 seconds to converge.

data, we note that the performance of the parallelization depends upon sample size as well as hardware configuration.

B.5 Supplementary Tables

Table B.1: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_3(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = 1$, $\beta_4(t) = \sin(3\pi t/4)$, and $\beta_5(t) = 1$. Effects $\beta_1(t)$ and $\beta_2(t)$ were expanded with 5 control points (knots) and the other 3 effects were expanded with 7 control points. Binary covariates for $\beta_1(t)$ and $\beta_2(t)$ had their frequency of being one varying uniformly from 0.8 to 0.9, while covariates for $\beta_3(t)$, $\beta_4(t)$ and $\beta_5(t)$ had their frequency of being one varying uniformly from 0.4 to 0.5.

method	size	IMSE	bias	variance
		Panel A: $\beta_1(t)$		
ProxiN	1000	7.16	1.12	5.90
	5000	1.34	0.08	1.33
	10000	0.67	0.13	0.65
NaiveN	1000	4970.70	16.53	4697.48
	5000	166.96	3.47	154.94
	10000	24.34	0.47	24.12
QuasiN	1000	565324.26	20.01	564923.92
	5000	4394.66	30.34	3473.92
	10000	7938.16	53.74	5050.27
		Panel B: $\beta_3(t)$		
ProxiN	1000	4.55	0.60	4.19
	5000	0.89	0.31	0.79
	10000	0.31	0.12	0.29
NaiveN	1000	343.03	3.05	333.73
	5000	6.29	0.57	5.96
	10000	4.76	0.50	4.51
QuasiN	1000	3136357.39	340.91	3020134.58
	5000	37874.74	75.48	32177.72
	10000	51506.41	111.61	39049.38

Table B.2: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_2(t)$ and $\hat{\beta}_4(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = 1$, $\beta_4(t) = \sin(3\pi t/4)$, and $\beta_5(t) = 1$. Effects $\beta_1(t)$ and $\beta_2(t)$ were expanded with 5 control points (knots) and the other 3 effects were expanded with 7 control points. Binary covariates for $\beta_1(t)$ and $\beta_2(t)$ had their frequency of being one varying uniformly from 0.8 to 0.9, while covariates for $\beta_3(t)$, $\beta_4(t)$ and $\beta_5(t)$ had their frequency of being one varying uniformly from 0.4 to 0.5.

method	size	IMSE	bias	variance
Panel A: $\beta_2(t)$				
ProxiN	1000	7.59	1.01	6.56
	5000	1.75	0.36	1.62
	10000	0.83	0.33	0.72
NaiveN	1000	15298.91	20.16	14892.65
	5000	2053.19	0.95	2052.29
	10000	147.46	1.95	143.67
QuasiN	1000	825978.27	122.77	810904.83
	5000	3629.98	30.64	2691.19
	10000	7695.80	64.09	3587.91
Panel B: $\beta_4(t)$				
ProxiN	1000	4.25	0.20	4.21
	5000	0.55	0.20	0.51
	10000	0.24	0.21	0.20
NaiveN	1000	169.73	0.91	168.89
	5000	1.78	0.30	1.69
	10000	1.55	0.34	1.44
QuasiN	1000	2680662.28	348.07	2559512.55
	5000	15715.84	56.26	12550.28
	10000	27001.54	95.26	17927.10

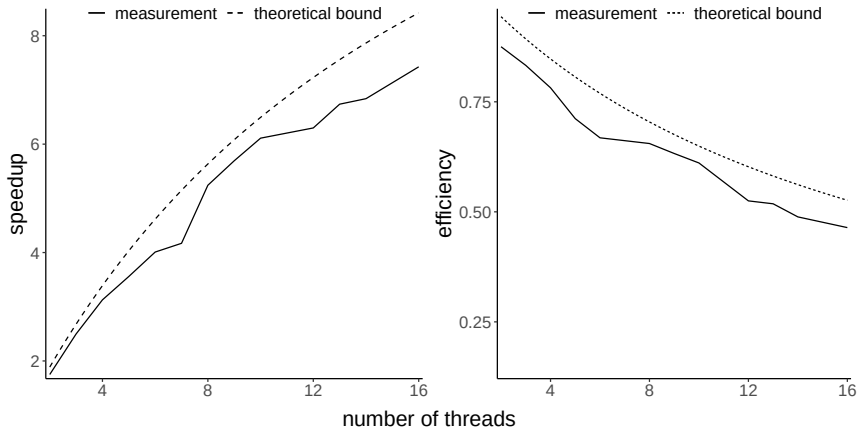


Figure B.3: Speedup and efficiency of the parallelized proximal Newton algorithm. Experiments were conducted using the SEER prostate cancer data on an Intel® Xeon® Gold 6254 quad-processor with max frequency 4 GHz and RAM 576 GB. The runtime was taken as an average of the duration of 10 runs with a fixed number of threads. The speedup and efficiency were defined in the caption of Figure B.2. In this case, the fraction of parallelizable code δ was around 94%. For reference, the serial proximal Newton algorithm took 71.35 seconds to converge.

Table B.3: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_{11}(t)$ and $\hat{\beta}_{12}(t)$ (corresponding to the first cause of failure) using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated. True values were $\beta_{11}(t) = 1$, $\beta_{12}(t) = \sin(3\pi t/4)$, $\beta_{13}(t) = -1$, $\beta_{14}(t) = -1$, and $\beta_{15}(t) = 1$. The effect $\beta_{11}(t)$ was expanded with 5 control points (knots) and the other 4 effects were expanded with 7 control points.

method	size	IMSE	bias	variance
Panel A: $\beta_{11}(t)$				
ProxiN	1000	1.88	0.06	1.88
	5000	0.60	0.08	0.60
	10000	0.44	0.07	0.43
NaiveN	1000	50.68	0.22	50.63
	5000	4.00	0.16	3.97
	10000	2.52	0.34	2.41
QuasiN	1000	10667.07	99.70	726.24
	5000	13967.36	113.29	1132.96
	10000	4966.15	68.71	245.43
Panel B: $\beta_{12}(t)$				
ProxiN	1000	2.96	0.31	2.86
	5000	1.05	0.27	0.98
	10000	0.66	0.19	0.63
NaiveN	1000	18046.24	29.96	17148.45
	5000	234.14	0.30	234.05
	10000	154.32	1.36	152.47
QuasiN	1000	29836.08	171.26	504.90
	5000	22516.75	145.01	1488.61
	10000	607.72	23.48	56.57

Table B.4: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_{21}(t)$ and $\hat{\beta}_{22}(t)$ (corresponding to the second cause of failure) using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated. True values were $\beta_{21}(t) = -1$, $\beta_{22}(t) = \cos(3\pi t/4)$, $\beta_{23}(t) = 1$, $\beta_{24}(t) = 1$, and $\beta_{25}(t) = -1$. The effect $\beta_{21}(t)$ was expanded with 5 control points (knots) and the other 4 effects were expanded with 7 control points.

method	size	IMSE	bias	variance
Panel A: $\beta_{21}(t)$				
ProxiN	1000	2.66	0.09	2.65
	5000	0.82	0.22	0.78
	10000	0.57	0.04	0.57
NaiveN	1000	210.96	2.62	204.10
	5000	11.50	0.11	11.49
	10000	9.63	0.40	9.47
QuasiN	1000	26201.67	160.88	320.23
	5000	26077.32	160.62	279.39
	10000	26457.63	161.71	309.09
Panel B: $\beta_{22}(t)$				
ProxiN	1000	4.41	0.41	4.24
	5000	1.12	0.31	1.02
	10000	0.73	0.28	0.65
NaiveN	1000	344296.28	85.31	337018.53
	5000	2242.58	9.38	2154.56
	10000	736.99	3.15	727.08
QuasiN	1000	816.01	19.09	451.39
	5000	832.45	27.32	86.07
	10000	1506.81	36.60	167.22

Table B.5: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated each with 1,000 observations. A fixed number of $K = 5$ knots were used for model fitting. The column 'censoring' indicates different uniform distributions of censoring times. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = -1$, $\beta_4(t) = -1$, and $\beta_5(t) = 1$.

method	censoring	IMSE	bias	variance
	Panel A: $\beta_1(t)$			
ProxiN	Uniform(0,3)	3.60	0.24	3.55
	Uniform(0.5,3)	2.82	0.15	2.80
	Uniform(1,3)	2.20	0.10	2.19
	Uniform(1.5,3)	1.63	0.18	1.59
NaiveN	Uniform(0,3)	35.82	0.99	34.84
	Uniform(0.5,3)	18.49	0.58	18.15
	Uniform(1,3)	17.95	0.68	17.48
	Uniform(1.5,3)	7.02	0.46	6.81
QuasiN	Uniform(0,3)	6772.12	69.37	1960.34
	Uniform(0.5,3)	6447.40	67.50	1891.83
	Uniform(1,3)	5843.91	65.85	1507.36
	Uniform(1.5,3)	5504.39	66.36	1100.59
	Panel B: $\beta_2(t)$			
ProxiN	Uniform(0,3)	1.82	0.28	1.74
	Uniform(0.5,3)	1.72	0.27	1.65
	Uniform(1,3)	1.39	0.24	1.33
	Uniform(1.5,3)	0.89	0.23	0.84
NaiveN	Uniform(0,3)	20.94	1.41	18.95
	Uniform(0.5,3)	11.05	1.02	10.01
	Uniform(1,3)	8.06	0.95	7.16
	Uniform(1.5,3)	3.24	0.64	2.83
QuasiN	Uniform(0,3)	72892.15	237.68	16400.41
	Uniform(0.5,3)	73688.59	241.31	15460.03
	Uniform(1,3)	67488.24	234.54	12481.04
	Uniform(1.5,3)	61153.12	230.64	7957.28

Table B.6: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated each with 1,000 observations. A fixed number of $K = 5$ knots were used for model fitting. The column 'censoring' indicates different exponential distributions of censoring times. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = -1$, $\beta_4(t) = -1$, and $\beta_5(t) = 1$.

method	censoring	IMSE	bias	variance
	Panel A: $\beta_1(t)$			
ProxiN	Exponential(0.2)	0.24	0.22	0.19
	Exponential(0.5)	0.68	0.22	0.63
	Exponential(0.8)	2.89	0.15	2.87
	Exponential(1.0)	4.65	0.36	4.53
NaiveN	Exponential(0.2)	0.25	0.22	0.20
	Exponential(0.5)	40.00	0.63	39.60
	Exponential(0.8)	45.88	0.20	45.84
	Exponential(1.0)	153.62	0.44	153.42
QuasiN	Exponential(0.2)	10738.67	93.24	2045.25
	Exponential(0.5)	11491.22	84.32	4381.66
	Exponential(0.8)	10737.09	77.90	4669.03
	Exponential(1.0)	50168.07	66.31	45771.43
	Panel B: $\beta_2(t)$			
ProxiN	Exponential(0.2)	0.26	0.31	0.16
	Exponential(0.5)	0.41	0.35	0.29
	Exponential(0.8)	1.28	0.36	1.15
	Exponential(1.0)	2.18	0.45	1.98
NaiveN	Exponential(0.2)	0.31	0.32	0.21
	Exponential(0.5)	2.65	0.49	2.40
	Exponential(0.8)	16.34	0.91	15.51
	Exponential(1.0)	53.29	2.35	47.79
QuasiN	Exponential(0.2)	42096.75	190.56	5783.34
	Exponential(0.5)	67970.58	212.90	22646.26
	Exponential(0.8)	82922.33	231.75	29212.65
	Exponential(1.0)	617933.47	179.31	585780.50

Table B.7: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated each with 1,000 observations. A fixed number of $K = 5$ knots were used for model fitting. Censoring times were generated from an exponential distribution with a rate of 0.5. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = -1$, $\beta_4(t) = -1$, and $\beta_5(t) = 1$.

method	size	IMSE	bias	variance
Panel A: $\beta_1(t)$				
ProxiN	100,000	0.074	0.27	0.00053
	1,000,000	0.074	0.27	0.00023
Panel B: $\beta_2(t)$				
ProxiN	100,000	0.15	0.39	0.00074
	1,000,000	0.15	0.39	0.00051

Table B.8: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes, where $t \in (0, 2]$. In each scenario, 100 data replicates were generated and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = -1$, $\beta_4(t) = -1$, and $\beta_5(t) = 1$.

method	size	IMSE	bias	variance
Panel A: $\beta_1(t)$				
ProxiN	1000	0.23	0.04	0.23
	5000	0.03	0.03	0.03
	10000	0.02	0.03	0.02
NaiveN	1000	0.97	0.09	0.96
	5000	0.02	0.02	0.02
	10000	0.01	0.01	0.01
QuasiN	1000	6414.11	69.50	1584.47
	5000	4784.41	47.07	2569.14
	10000	3449.31	42.66	1629.44
Panel B: $\beta_2(t)$				
ProxiN	1000	0.19	0.06	0.19
	5000	0.04	0.07	0.03
	10000	0.02	0.07	0.01
NaiveN	1000	0.54	0.17	0.51
	5000	0.05	0.09	0.04
	10000	0.02	0.07	0.02
QuasiN	1000	66114.46	227.58	14323.58
	5000	46499.04	122.28	31546.56
	10000	31094.54	117.32	17330.83

Table B.9: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes, where $t \in (2, 3]$. In each scenario, 100 data replicates were generated and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_1(t) = 1$, $\beta_2(t) = \sin(3\pi t/4)$, $\beta_3(t) = -1$, $\beta_4(t) = -1$, and $\beta_5(t) = 1$.

method	size	IMSE	bias	variance
Panel A: $\beta_1(t)$				
ProxiN	1000	11.13	0.43	10.94
	5000	0.75	0.01	0.75
	10000	0.42	0.06	0.41
NaiveN	1000	112.19	1.76	109.08
	5000	0.79	0.04	0.79
	10000	0.42	0.09	0.42
QuasiN	1000	8152.12	72.22	2936.10
	5000	6371.02	33.61	5241.41
	10000	5029.56	47.96	2729.46
Panel B: $\beta_2(t)$				
ProxiN	1000	5.49	0.51	5.23
	5000	0.50	0.35	0.38
	10000	0.38	0.38	0.24
NaiveN	1000	5.60	2.48	59.44
	5000	0.68	0.36	0.55
	10000	0.35	0.33	0.24
QuasiN	1000	90108.10	265.80	19459.02
	5000	38950.91	78.33	32815.05
	10000	18420.84	85.59	11095.19

Table B.10: Integrated mean squared error (IMSE), average bias, and average variance of estimates $\hat{\beta}_{21}(t)$ and $\hat{\beta}_{22}(t)$ (corresponding to the second cause of failure) using the proximal Newton (ProxiN), naive Newton (NaiveN), and quasi-Newton (QuasiN) methods with varying sample sizes. In each scenario, 100 data replicates were generated, and a fixed number of $K = 5$ knots were used for model fitting. True values were $\beta_{21}(t) = -1$, $\beta_{22}(t) = \cos(3\pi t/4)$, $\beta_{23}(t) = 1$, $\beta_{24}(t) = 1$, and $\beta_{25}(t) = -1$.

method	size	IMSE	bias	variance
Panel A: $\beta_{21}(t)$				
ProxiN	1000	2.74	0.08	2.74
	5000	0.87	0.25	0.81
	10000	0.59	0.05	0.59
NaiveN	1000	184.50	4.13	167.44
	5000	11.17	0.04	11.17
	10000	7.77	0.34	7.66
QuasiN	1000	228079.67	52.04	225371.79
	5000	7137.96	58.58	3706.53
	10000	6077.03	64.58	1906.45
Panel B: $\beta_{22}(t)$				
ProxiN	1000	3.32	0.62	2.94
	5000	1.22	0.49	0.98
	10000	0.95	0.51	0.69
NaiveN	1000	1283.69	4.58	1262.67
	5000	20.77	0.64	20.36
	10000	68.89	0.56	68.57
QuasiN	1000	971229.96	100.36	961157.95
	5000	14782.05	83.76	7765.94
	10000	9854.26	81.96	3137.26

APPENDIX C

Supplementary Material for Chapter IV

C.1 Gradient and Hessian of $\ell_{gj}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$

For $g = 1, \ldots, G$, $i = 1, \ldots, n_g$, and $j = 1, \ldots, m$, we define

$$S_{gij}^{(u)}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \coloneqq \sum_{r \in R_g(X_{gi})} \exp\{\mathbf{L}_{gr}^\top(X_{gi})\boldsymbol{\gamma}_j + \mathbf{W}_{gr}^\top \boldsymbol{\theta}_j\} \begin{bmatrix} \mathbf{L}_{gr}(X_{gi}) \\ \mathbf{W}_{gr} \end{bmatrix}^{\odot u}, \quad u = 0, 1, 2,$$

where $\mathbf{L}_{gr}(X_{gi}) \coloneqq \mathbf{Z}_{gr} \otimes \breve{\mathbf{B}}(\breve{X}_{gr}) \otimes \mathbf{B}(X_{gi})$, and for a vector $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^{\odot 0} \coloneqq 1$, $\mathbf{v}^{\odot 1} \coloneqq \mathbf{v}$, and $\mathbf{v}^{\odot 2} \coloneqq \mathbf{v}\mathbf{v}^\top$. The gradient $\dot{\ell}_{gj}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ and Hessian $\ddot{\ell}_{gj}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ of $\ell_{gj}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j)$ are hence given by

$$\dot{\ell}_{gj}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j) = \sum_{i=1}^{n_g} \Delta_{gij} \left\{ \begin{bmatrix} \mathbf{L}_{gi}(X_{gi}) \\ \mathbf{W}_{gi} \end{bmatrix} - \mathbf{U}_{gij}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \right\},$$

$$\ddot{\ell}_{gj}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j) = -\sum_{i=1}^{n_g} \Delta_{gij} \mathbf{V}_{gij}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}),$$

in which

$$\mathbf{U}_{gij}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \coloneqq \frac{S_{gij}^{(1)}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})}{S_{gij}^{(0)}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})}, \quad \mathbf{V}_{gij}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}) \coloneqq \frac{S_{gij}^{(2)}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})}{S_{gij}^{(0)}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi})} - \mathbf{U}_{gij}^{\odot 2}(\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j, X_{gi}).$$

C.2 Proof of Proposition IV.1

Proposition C.1. *Under* $H_0^{(t)} : \mathbf{C}^{(t)}\mathrm{vec}(\boldsymbol{\gamma}_{jl}^\top) = \mathbf{0}$, *the test statistic*

$$\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}^\top \{\mathbf{C}^{(t)}\}^\top \left[\mathbf{C}^{(t)} \boldsymbol{\Omega}_{jl} \{\mathbf{C}^{(t)}\}^\top\right]^{-1} \mathbf{C}^{(t)} \{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$$

asymptotically follows a distribution characterized by

$$\sum_{u=1}^{K\check{K}\times K\check{K}} \mu_u G_u^2,$$

where G_u's are independent standard normal random variables, and μ_u's are the possibly identical eigenvalues of the matrix product of $[\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top]^{-1}$ and the variance of $\mathbf{C}^{(t)}\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$.

Proof. Let $\mathbf{M} \coloneqq [\mathbf{C}^{(t)}\boldsymbol{\Omega}_{jl}\{\mathbf{C}^{(t)}\}^\top]^{-1}$, let $\boldsymbol{\Sigma}$ denote the variance of $\mathbf{x} \coloneqq \mathbf{C}^{(t)}\{\mathrm{vec}(\tilde{\boldsymbol{\gamma}}_{jl}^\top) - \tilde{\mathbf{b}}_{jl}\}$, and let $Q \coloneqq \mathbf{x}^\top\mathbf{M}\mathbf{x}$ denote the Wald test statistic. Since $\boldsymbol{\Sigma}$ is orthogonally diagonalizable, there exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^\top = \boldsymbol{\Psi}$, with $\boldsymbol{\Psi}$ being a diagonal matrix of positive eigenvalues of $\boldsymbol{\Sigma}$. Let $\mathbf{R} \coloneqq \boldsymbol{\Psi}^{-1/2}\mathbf{P}$, a nonsingular matrix. Then $\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^\top = \mathbf{I}$. Since $(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}$ is symmetric and orthogonally diagonalizable, there exists another orthogonal matrix $\mathbf{T}$ such that $\mathbf{T}(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{T}^\top = \boldsymbol{\Phi}$ is a diagonal matrix sharing the same eigenvalues $\mu_1, \ldots, \mu_{K\check{K}\times K\check{K}}$ as those of $(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}$. Let $\mathbf{z} \coloneqq \mathbf{T}\mathbf{R}\mathbf{x}$. Then under the null $H_0^{(t)}$, we have $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since $\mathbf{T}\mathbf{R}$ is nonsingular, $\mathbf{x} = \mathbf{R}^{-1}\mathbf{T}^\top\mathbf{z}$. It follows that $Q = \mathbf{z}^\top\boldsymbol{\Phi}\mathbf{z} = \sum_{u=1}^{K\check{K}\times K\check{K}} \mu_u G_u^2$, where G_u's independently follow the standard normal distribution. Observe that

$$(\mathbf{R}^\top\mathbf{T}^\top)^{-1}\mathbf{M}\boldsymbol{\Sigma}\mathbf{R}^\top\mathbf{T}^\top = \mathbf{T}(\mathbf{R}^\top)^{-1}\mathbf{M}\boldsymbol{\Sigma}\mathbf{R}^\top\mathbf{T}^\top = \mathbf{T}(\mathbf{R}^\top)^{-1}\mathbf{M}\mathbf{R}^{-1}\mathbf{T}^\top = \boldsymbol{\Phi}.$$

This implies that $\mathbf{M}\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$ have the same set of eigenvalues (since the mapping $\mathbf{A} \mapsto \mathbf{B}^{-1}\mathbf{A}\mathbf{B}$ preserves eigenvalues). ■

C.3 Proof of Proposition IV.2

Proposition C.2. *Let $\hat{\lambda}_{0jg}(\cdot)$ be the estimated baseline hazard function derived from the unpenalized bivariate varying coefficient model. Let*

$$\tilde{M}_{jgi} \coloneqq \Delta_{jgi} - \exp(\mathbf{W}_{gi}^\top\tilde{\boldsymbol{\theta}}_j^{-f}) \int_0^{X_{gi}} \exp\left\{\mathbf{Z}_{gi}^\top\tilde{\boldsymbol{\beta}}_j^{-f}(t, \check{X}_{gi})\right\} \hat{\lambda}_{0jg}(t)\,\mathrm{d}t$$

be the martingale residual for subject i in the gth stratum, where $\tilde{\boldsymbol{\beta}}_j^{-f}(\cdot,\cdot)$ and $\tilde{\boldsymbol{\theta}}_j^{-f}$ are the penalized estimates from the corresponding fold f to which subject i in the gth stratum belongs. Then the deviance residual for subject i in the gth stratum with respect to the jth failure type is written as

$$d_{jgi} \coloneqq \operatorname{sign}(\tilde{M}_{jgi})\sqrt{-2\left[\Delta_{jgi}\left\{\mathbf{Z}_{gi}^\top\tilde{\boldsymbol{\beta}}_j^{-f}(X_{gi},\breve{X}_{gi})+\mathbf{W}_{gi}^\top\tilde{\boldsymbol{\theta}}_j^{-f}+\log\int_0^{X_{gi}}\hat{\lambda}_{0jg}(t)\,\mathrm{d}t\right\}+\tilde{M}_{jgi}\right]}.$$

Proof. Given estimates $\hat{\boldsymbol{\theta}}_j$, $\hat{\boldsymbol{\beta}}_j(\cdot,\cdot)$ for the bivariate varying coefficient model (1), the martingale residuals can be defined as

$$\hat{M}_{jgi} \coloneqq \hat{M}_{jgi}(\infty,\breve{X}_{gi}) = \Delta_{jgi} - \exp(\mathbf{W}_{gi}^\top\hat{\boldsymbol{\theta}}_j)\int_0^{X_{gi}}\exp\left\{\mathbf{Z}_{gi}^\top\hat{\boldsymbol{\beta}}_j(t,\breve{X}_{gi})\right\}\hat{\lambda}_{0jg}(t)\,\mathrm{d}t,$$

where the baseline hazard estimates $\hat{\lambda}_{0jg}(\cdot)$ are determined via the Breslow estimator. Further, the log-likelihood with respect to the jth failure type can be written as

$$\begin{aligned}
&\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{\Delta_{jgi}\log\lambda_{jgi}(X_{gi}\mid\mathbf{Z}_{gi},\mathbf{W}_{gi},\breve{X}_{gi})+\log S_{jgi}(X_{gi}\mid\mathbf{Z}_{gi},\mathbf{W}_{gi},\breve{X}_{gi})\}\\
&=\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Bigg[\Delta_{jgi}\{\mathbf{Z}_{gi}^\top\boldsymbol{\beta}_j(X_{gi},\breve{X}_{gi})+\mathbf{W}_{gi}^\top\boldsymbol{\theta}_j+\log\lambda_{0jg}(X_{gi})\\
&\qquad-\int_0^{X_{gi}}\exp\{\mathbf{Z}_{gi}^\top\boldsymbol{\beta}_j(t,\breve{X}_{gi})+\mathbf{W}_{gi}^\top\boldsymbol{\theta}_j\}\lambda_{0jg}(t)\,\mathrm{d}t\Bigg],
\end{aligned}$$

where $S_{jgi}(t\mid\mathbf{Z}_{gi},\mathbf{W}_{gi},\breve{X}_{gi})$ is the corresponding survivor function. Assuming that the baseline hazard $\lambda_{0jg}(\cdot)$ is known, we have the deviance D written as

$$\begin{aligned}
D = 2\sup_{\boldsymbol{\beta}_{jgi},\boldsymbol{\theta}_{jgi}}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Bigg\{&\Delta_{jgi}[\mathbf{Z}_{gi}^\top\{\boldsymbol{\beta}_{jgi}-\hat{\boldsymbol{\beta}}_j(X_{gi},\breve{X}_{gi})\}+\mathbf{W}_{gi}^\top(\boldsymbol{\theta}_{jgi}-\hat{\boldsymbol{\theta}}_j)]\\
&-\int_0^{X_{gi}}\left[\exp(\mathbf{Z}_{gi}^\top\boldsymbol{\beta}_{jgi}+\mathbf{W}_{gi}^\top\boldsymbol{\theta}_{jgi})-\exp\{\mathbf{Z}_{gi}^\top\hat{\boldsymbol{\beta}}_j(t,\breve{X}_{gi})+\mathbf{W}_{gi}^\top\hat{\boldsymbol{\theta}}_j\}\right]\lambda_{0jg}(t)\,\mathrm{d}t\Bigg\},
\end{aligned}$$

where $\boldsymbol{\beta}_{jgi}$ and $\boldsymbol{\theta}_{jgi}$ are subject-cause-specific estimates allowed in a saturated model. Now, we have the first order condition

$$\Delta_{jgi} = \exp(\mathbf{Z}_{gi}^\top\boldsymbol{\beta}_{jgi}+\mathbf{W}_{gi}^\top\boldsymbol{\theta}_{jgi})\int_0^{X_{gi}}\lambda_{0jg}(t)\,\mathrm{d}t,\quad g=1,\ldots,G,\ i=1,\ldots,n_g.$$

With this condition, the deviance D reduces to

$$\begin{aligned} D &= -2\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left\{\Delta_{jgi}\log\frac{\exp\{\mathbf{Z}_{gi}^{\top}\hat{\boldsymbol{\beta}}_j(X_{gi},\breve{X}_{gi})+\mathbf{W}_{gi}^{\top}\hat{\boldsymbol{\theta}}_j\}\int_0^{X_{gi}}\lambda_{0jg}(t)\,\mathrm{d}t}{\Delta_{jgi}}+\tilde{M}_{jgi}\right\} \\ &= -2\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left[\Delta_{jgi}\left\{\mathbf{Z}_{gi}^{\top}\hat{\boldsymbol{\beta}}_j(X_{gi},\breve{X}_{gi})+\mathbf{W}_{gi}^{\top}\hat{\boldsymbol{\theta}}_j+\log\int_0^{X_{gi}}\lambda_{0jg}(t)\,\mathrm{d}t\right\}+\tilde{M}_{jgi}\right], \end{aligned}$$

where

$$\tilde{M}_{jgi}\coloneqq\tilde{M}_{jgi}(\infty,\breve{X}_{gi})=\Delta_{jgi}-\exp(\mathbf{W}_{gi}^{\top}\hat{\boldsymbol{\theta}}_j)\int_0^{X_{gi}}\exp\left\{\mathbf{Z}_{gi}^{\top}\hat{\boldsymbol{\beta}}_j(t,\breve{X}_{gi})\right\}\lambda_{0jg}(t)\,\mathrm{d}t$$

is the martingale residual with known baseline hazard $\lambda_{0jg}(\cdot)$. Then the deviance residual d_{jgi} for subject i in the gth stratum with respect to the jth failure type can be written as

$$d_{jgi}=\mathrm{sign}(\hat{M}_{jgi})\sqrt{-2\left[\Delta_{jgi}\left\{\mathbf{Z}_{gi}^{\top}\hat{\boldsymbol{\beta}}_j(X_{gi},\breve{X}_{gi})+\mathbf{W}_{gi}^{\top}\hat{\boldsymbol{\theta}}_j+\log\int_0^{X_{gi}}\hat{\lambda}_{0jg}(t)\,\mathrm{d}t\right\}+\hat{M}_{jgi}\right]},$$

where $\hat{M}_{jgi}$ is the martingale residual $\tilde{M}_{jgi}$ with $\lambda_{0jg}(\cdot)$ replaced by $\hat{\lambda}_{0jg}(\cdot)$. ■

C.4 Supplementary Figure

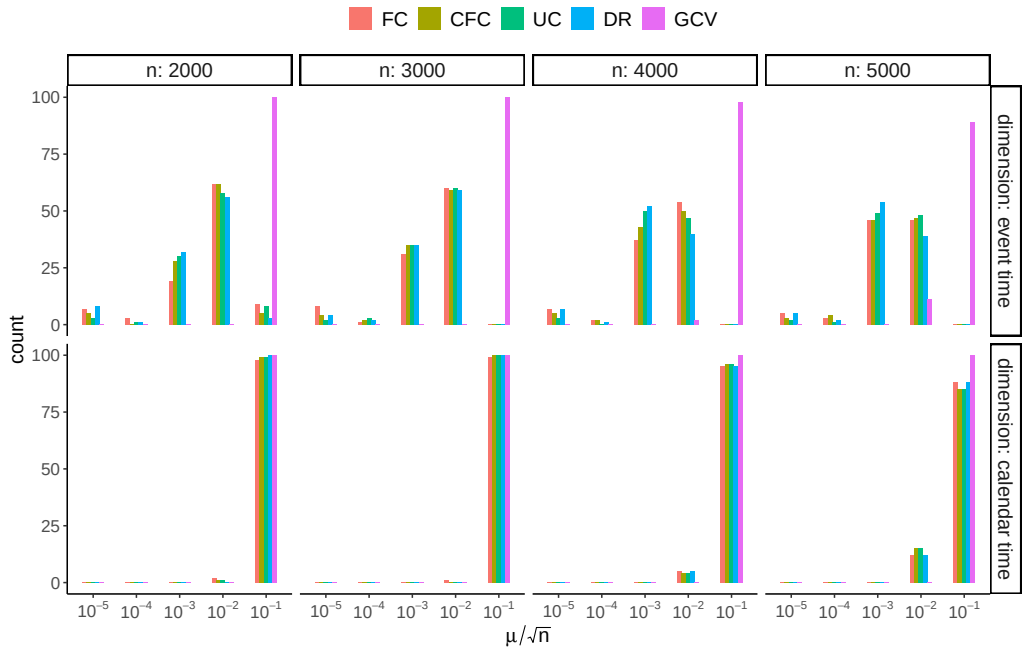


Figure C.1: A comparison of the distribution of selected tuning parameters for five cross-validation methods: fold-constrained (FC), complementary fold-constrained (CFC), and fold-unconstrained (UC) cross-validated partial likelihood, cross-validated deviance residuals (DR), and generalized cross-validation (GCV). In each scenario, 100 training and validation data replicates were generated independently. A 5-by-5 grid of tuning parameters was formed such that $\mu/\sqrt{n}$ (with n denoting sample size) and $\breve{\mu}/\sqrt{n}$ varied from 10^{-5} to 10^{-1}. Each cross-validation method was applied to a training data replicate to determine the optimal tuning parameters. True values were $\beta_1(t, \breve{x}) = \sin(3\pi t/4)\exp(-0.5\breve{x})$ and $\beta_2(t, \breve{x}) = 1$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] G. M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), pages 483–485. ACM, 1967.

[2] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[3] A. S. Ash, S. F. Fienberg, T. A. Louis, S.-L. T. Normand, T. A. Stukel, and J. Utts. Statistical Issues in Assessing Hospital Performance. Commissioned by the Committee of Presidents of Statistical Societies. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf, 2012. Accessed: 2020-08-19.

[4] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, second edition, 2015.

[5] S. Baulies, L. Belin, P. Mallon, C. Senechal, J. Pierga, P. Cottu, M. Sablin, X. Sastre, B. Asselain, R. Rouzier, et al. Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy. *British Journal of Cancer*, 113(1):30–36, 2015.

[6] C. A. Bellera, G. MacGrogan, M. Debled, C. T. de Lara, V. Brouste, and S. Mathoulin-Pélissier. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*, 10(1):1–12, 2010.

[7] J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher. Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6):956–971, 2009.

[8] D. D. Boos. On generalized score tests. *The American Statistician*, 46(4):327–333, 1992.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, second edition, 2004.

[10] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253, 2011.

[11] A. F. Brouwer, K. He, S. B. Chinn, A. M. Mondul, C. H. Chapman, M. D. Ryser, M. Banerjee, M. C. Eisenberg, R. Meza, and J. M. G. Taylor. Time-varying survival effects for squamous cell carcinomas at oropharyngeal and nonoropharyngeal head and neck sites in the United States, 1973-2015. *Cancer*, 126(23):5137–5146, 2020.

[12] H. Casanova, A. Legrand, and Y. Robert. *Parallel Algorithms*. CRC Press, 2008.

[13] D. Castner. Management of patients on hemodialysis before, during, and after hospitalization: challenges and suggestions for improvements. *Nephrology Nursing Journal*, 38(4):319–330, 2011.

[14] Centers for Disease Control and Prevention. Chronic Kidney Disease in the United States, 2019. https://www.cdc.gov/kidneydisease/publications-resources/2019-national-facts.html, 2019. Accessed: 2020-8-19.

[15] Centers for Medicare & Medicaid Services. Medicare Beneficiaries at a Glance. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Beneficiary-Snapshot/Downloads/Bene_Snaphot.pdf, 2021. Accessed: 2021-12-24.

[16] Centers for Medicare and Medicaid Services. ESRD Quality Incentive Program. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/ESRDQIP, 2020. Accessed: 2021-12-24.

[17] K. E. Chan, J. M. Lazarus, R. L. Wingard, and R. M. Hakim. Association between repeat hospitalization and early intervention in dialysis patients following hospital discharge. *Kidney International*, 76(3):331–341, 2009.

[18] CMS. Care Compare–Dialysis Facilities. https://www.medicare.gov/care-compare/, 2021. Accessed: 2021-12-26.

[19] CMS. ESRD Quality Incentive Program. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/ESRDQIP, 2021. Accessed: 2021-12-26.

[20] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972.

[21] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[22] B. Dai and P. Breheny. Cross validation approaches for penalized Cox regression. https://arxiv.org/abs/1905.10432, 2019. Accessed: 2022-02-26.

[23] R. B. Davies. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):323–333, 1980.

[24] C. de Boor. *A Practical Guide to Splines*. Springer, Revised edition, 2001.

[25] D. J. de Jager, D. C. Grootendorst, K. J. Jager, P. C. van Dijk, L. M. Tomas, D. Ansell, F. Collart, P. Finne, J. G. Heaf, J. De Meester, J. F. M. Wetzels, F. R. Rosendaal, and F. W. Dekker. Cardiovascular and noncardiovascular mortality among patients starting dialysis. *JAMA*, 302(16):1782–1789, 2009.

[26] R. de Mutsert, M. B. Snijder, F. van der Sman-de Beer, J. C. Seidell, E. W. Boeschoten, R. T. Krediet, J. M. Dekker, J. P. Vandenbroucke, F. W. Dekker, et al. Association between body mass index and mortality is similar in the hemodialysis population and the general population at high age and equal duration of follow-up. *Journal of the American Society of Nephrology*, 18(3):967–974, 2007.

[27] F. W. Dekker, R. de Mutsert, P. C. Van Dijk, C. Zoccali, and K. J. Jager. Survival analysis: time-dependent effects and time-varying risk factors. *Kidney International*, 74(8):994–997, 2008.

[28] E. Demidenko. *Mixed Models: Theory and Applications with R*. John Wiley & Sons, 2013.

[29] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

[30] T.-N. Do and F. Poulet. Parallel multiclass logistic regression for classifying large scale image datasets. In *Advanced Computational Methods for Knowledge Engineering*, pages 255–266. Springer, 2015.

[31] D. Eddelbuettel. CRAN Task View: High-Performance and Parallel Computing with R. https://cran.r-project.org/web/views/HighPerformanceComputing.html, 2021. Accessed: 2021-01-26.

[32] D. Eddelbuettel and J. J. Balamuta. Extending R with C++: A brief introduction to Rcpp. *The American Statistician*, 72(1):28–36, 2018.

[33] D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[34] D. Eddelbuettel and C. Sanderson. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, 71:1054–1063, 2014.

[35] P. H. Eilers and B. D. Marx. *Practical Smoothing: The Joys of P-splines*. Cambridge University Press, 2021.

[36] J. P. Estes, Y. Chen, D. Şentürk, C. M. Rhee, E. Kürüm, A. S. You, E. Streja, K. Kalantar-Zadeh, and D. V. Nguyen. Profiling dialysis facilities for adverse recurrent events. *Statistics in Medicine*, 39(9):1374–1389, 2020.

[37] J. P. Estes, D. V. Nguyen, Y. Chen, L. S. Dalrymple, C. M. Rhee, K. Kalantar-Zadeh, and D. Şentürk. Time-dynamic profiling with application to hospital readmission among patients on dialysis. *Biometrics*, 74(4):1383–1394, 2018.

[38] J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1):179, 2008.

[39] A. O. Finley and S. Banerjee. Bayesian spatially varying coefficient models in the spBayes R package. *Environmental Modelling & Software*, 125:104608, 2020.

[40] A. R. Gallant. *Nonlinear Statistical Models*, volume 310. John Wiley & Sons, 1987.

[41] A. E. Gelfand, H.-J. Kim, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

[42] A. A. Goldstein. *Constructive Real Analysis*. Harper & Row, 1967.

[43] R. J. Goudie, R. M. Turner, D. De Angelis, and A. Thomas. MultiBUGS: A parallel implementation of the BUGS modelling framework for faster Bayesian inference. *Journal of Statistical Software*, 95(7):1–20, 2020.

[44] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[45] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992.

[46] R. J. Gray. Spline-based tests in survival analysis. *Biometrics*, 50(3):640–652, 1994.

[47] S. Haneuse and K. H. Lee. Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circulation: Cardiovascular Quality and Outcomes*, 9(3):322–331, 2016.

[48] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55(4):757–779, 1993.

[49] K. He, J. D. Kalbfleisch, Y. Li, and Y. Li. Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Analysis*, 19(4):490–512, 2013.

[50] K. He, Y. Yang, Y. Li, J. Zhu, and Y. Li. Modeling time-varying effects with large-scale survival data: An efficient quasi-Newton approach. *Journal of Computational and Graphical Statistics*, 26(3):635–645, 2017.

[51] K. He, J. Zhu, J. Kang, and Y. Li. Stratified cox models with time-varying effects for national kidney transplant patients: A new block-wise steepest ascent method. *Biometrics*, 2021.

[52] J. Hester and D. Schmidt. bench: High Precision Timing of R Expressions. https://cran.r-project.org/package=bench, 2020. R package version 1.1.1.

[53] L. Horwitz, C. Partovain, Z. Lin, J. Herrin, J. Grady, M. Conover, J. Montague, C. Dillaway, K. Bartczak, J. Ross, S. Bernheim, E. Drye, and H. Krumholz. Hospital-wide all-cause risk-standardized readmission measure: DRAFT Measure Methodology Report. Submitted by Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (YNHHSC/CORE). https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/downloads/MMSHospital-WideAll-ConditionReadmissionRate.pdf, 2011. Accessed: 2020-08-19.

[54] R. Jyothi and P. Babu. Piano: A fast parallel iterative algorithm for multinomial and sparse multinomial logistic regression. https://arxiv.org/abs/2002.09133, 2020. Accessed: 2021-09-14.

[55] K. Kalantar-Zadeh. Causes and consequences of the reverse epidemiology of body mass index in dialysis patients. *Journal of Renal Nutrition*, 15(1):142–147, 2005.

[56] K. Kalantar-Zadeh, G. Block, M. H. Humphreys, and J. D. Kopple. Reverse epidemiology of cardiovascular risk factors in maintenance dialysis patients. *Kidney International*, 63(3):793–808, 2003.

[57] J. D. Kalbfleisch and K. He. Discussion on "Time-dynamic profiling with application to hospital readmission among patients on dialysis," by Jason P. Estes, Danh V. Nguyen, Yanjun Chen, Lorien S. Dalrymple, Connie M. Rhee, Kamyar Kalantar-Zadeh, and Damla Senturk. *Biometrics*, 74(4):1401–1403, 2018.

[58] J. D. Kalbfleisch and R. L. Prentice. Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

[59] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Second edition, 2002.

[60] J. D. Kalbfleisch and R. A. Wolfe. On monitoring outcomes of medical providers. *Statistics in Biosciences*, 5(2):286–302, 2013.

[61] Kidney Epidemiology and Cost Center. ESRD emergency department visits technical expert panel summary report. https://dialysisdata.org/sites/default/files/content/ESRD_Measures/ESRD_Emergency_Department_Visits_TEP_Summary_Report.pdf, 2016. Accessed: 2020-08-19.

[62] Kidney Epidemiology and Cost Center. Guide to the quarterly dialysis facility compare – Preview report for October 2020 release. https://dialysisdata.org/sites/default/files/content/DFC_Guide_October2020.pdf, 2020. Accessed: 2020-08-19.

[63] Kidney Epidemiology and Cost Center. Technical notes on the standardized mortality ratio for the Dialysis Facility Reports. https://dialysisdata.org/sites/default/files/content/SMR%20Documentation.pdf, 2020. Accessed: 2020-08-19.

[64] M. Kim and L. Wang. Generalized spatially varying coefficient models. *Journal of Computational and Graphical Statistics*, 30(1):1–10, 2021.

[65] M. Kim, L. Wang, and Y. Zhou. Spatially varying coefficient models with sign preservation of the coefficient functions. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–20, 2021.

[66] P. Lafaye de Micheaux. *CompQuadForm: Distribution Function of Quadratic Forms in Normal Variables*, 2017. R package version 1.4.3.

[67] M.-J. Lai and L. L. Schumaker. *Spline Functions on Triangulations*. Cambridge University Press, 2007.

[68] K. Lange. *Optimization*. Springer Science & Business Media, second edition, 2013.

[69] J. D. Lee, Y. Sun, and M. Saunders. Proximal Newton-type methods for convex optimization. *Advances in Neural Information Processing Systems*, 25:827–835, 2012.

[70] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[71] K. H. Lee, F. Dominici, D. Schrag, and S. Haneuse. Hierarchical models for semicompeting risks data with application to quality of end-of-life care for pancreatic cancer. *Journal of the American Statistical Association*, 111(515):1075–1095, 2016.

[72] K. H. Lee, S. Haneuse, D. Schrag, and F. Dominici. Bayesian semi-parametric analysis of semicompeting risks data: Investigating hospital readmission after a pancreatic cancer diagnosis. *Journal of the Royal Statistical Society: Series C*, 64(2):253–273, 2015.

[73] M. Lee, E. J. Feuer, and J. P. Fine. On the analysis of discrete time competing risks data. *Biometrics*, 74(4):1468–1481, 2018.

[74] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.

[75] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[76] F.-C. Lin and J. Zhu. Additive hazards regression and partial likelihood estimation for ecological monitoring data across space. *Statistics and Its Interface*, 5(2):195–206, 2012.

[77] Y. Liu and F. Guo. A Bayesian time-varying coefficient model for multitype recurrent events. *Journal of Computational and Graphical Statistics*, 29(2):383–395, 2020.

[78] C.-L. Lu, S. Wang, Z. Ji, Y. Wu, L. Xiong, X. Jiang, and L. Ohno-Machado. WebDISCO: a web service for distributed Cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219, 2015.

[79] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[80] G. McGee, J. Schildcrout, S.-L. Normand, and S. Haneuse. Outcome-dependent sampling in cluster-correlated data settings with application to hospital profiling. *Journal of the Royal Statistical Society: Series A*, 183(1):379–402, 2020.

[81] J. Mu. *Spatially Varying Coefficient Models: Theory and Methods*. PhD thesis, Iowa State University, 2019. https://lib.dr.iastate.edu/etd/17519.

[82] J. Mu, G. Wang, and L. Wang. Estimation and inference in spatially varying coefficient models. *Environmetrics*, 29(1):e2485, 2018.

[83] D. Murakami and D. A. Griffith. Spatially varying coefficient modeling for large datasets: Eliminating n from spatial regressions. *Spatial Statistics*, 30:39–64, 2019.

[84] National Cancer Institute. Overview of the SEER Program. https://seer.cancer.gov/about/overview.html, 2021. Accessed: 2021-12-24.

[85] X. Niu and H. R. Cho. Adjusting for baseline information in comparing the efficacy of treatments using bivariate varying-coefficient models. *Journal of Nonparametric Statistics*, 31(3):680–694, 2019.

[86] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[87] S.-L. T. Normand, M. E. Glickman, and C. A. Gatsonis. Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, 92(439):803–814, 1997.

[88] D. I. Ohlssen, L. D. Sharples, and D. J. Spiegelhalter. A hierarchical modelling framework for identifying unusual performance in health care providers. *Journal of the Royal Statistical Society: Series A*, 170(4):865–890, 2007.

[89] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, 1970.

[90] W. Pan. On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906, 2001.

[91] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[92] H. Peng, D. Liang, and C. Choi. Evaluating parallel logistic regression models. In *2013 IEEE International Conference on Big Data*, pages 119–126. IEEE, 2013.

[93] A. Perperoglou, S. le Cessie, and H. C. van Houwelingen. A fast routine for fitting Cox models with time varying effects of the covariates. *Computer Methods and Programs in Biomedicine*, 81(2):154–161, 2006.

[94] M. Pietrosanu, H. Shu, B. Jiang, L. Kong, G. Heo, Q. He, J. Gilmore, and H. Zhu. Estimation for the bivariate quantile varying coefficient model with application to diffusion tensor imaging data analysis. *Biostatistics*, 2021.

[95] R. L. Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):1033–1048, 1988.

[96] R. L. Prentice and L. A. Gloeckler. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34(1):57–67, 1978.

[97] R. L. Prentice and J. D. Kalbfleisch. Mixed discrete and continuous Cox regression model. *Lifetime Data Analysis*, 9(2):195–210, 2003.

[98] R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson Jr, N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554, 1978.

[99] R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org, 2020.

[100] J. B. Reilly, L. M. Marcotte, J. S. Berns, and J. A. Shea. Handoff communication between hospital and outpatient dialysis units at patient discharge: a qualitative study. *The Joint Commission Journal on Quality and Patient Safety*, 39(2):70–76, 2013.

[101] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[102] A. Rotnitzky and N. P. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77(3):485–497, 1990.

[103] L. M. Sangalli, J. O. Ramsay, and T. O. Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703, 2013.

[104] R. Saran, B. Robinson, K. C. Abbott, L. Y. Agodoa, J. Bragg-Gresham, R. Balkrishnan, N. Bhave, et al. US Renal Data System 2018 Annual Data Report: Epidemiology of Kidney Disease in the United States. *American Journal of Kidney Diseases*, 73(3):A7–A8, 2019.

[105] E. L. Schreiber, R. E. Korf, and M. D. Moffitt. Optimal multi-way number partitioning. *Journal of the ACM*, 65(4):1–61, 2018.

[106] L. Schumaker. *Spline Functions: Basic Theory*. Cambridge University Press, third edition, 2007.

[107] L. L. Schumaker. *Spline Functions: Computational Methods*. Society for Industrial and Applied Mathematics, 2015.

[108] N. Serban. A space-time varying coefficient model: The equity of service accessibility. *Annals of Applied Statistics*, pages 2024–2051, 2011.

[109] V. B. Shahinian, X. Zhang, A. M. Tilea, K. He, D. E. Schaubel, W. Wu, R. Pisoni, B. Robinson, R. Saran, and K. J. Woodside. Surgeon characteristics and dialysis vascular access outcomes in the United States: A retrospective cohort study. *American Journal of Kidney Diseases*, 75(2):158–166, 2020.

[110] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[111] D. J. Spiegelhalter. Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24(8):1185–1202, 2005.

[112] R. J. Sun and J. D. Kalbfleisch. A risk-adjusted O–E CUSUM with monitoring bands for monitoring medical outcomes. *Biometrics*, 69(1):62–69, 2013.

[113] Y. Sun, H. Yan, W. Zhang, and Z. Lu. A semiparametric spatial dynamic model. *Annals of Statistics*, 42(2):700–727, 2014.

[114] Surveillance, Epidemiology, and End Results Program. Incidence - SEER 9 Regs Research Data, Nov 2017 Sub (1973-2015) <Katrina/Rita Population Adjustment>. https://seer.cancer.gov/data-software/documentation/seerstat/nov2017, 2017. Accessed: 2021-1-26.

[115] Surveillance, Epidemiology, and End Results Program. Incidence - SEER Research Data, 18 Registries, Nov 2019 Sub (2000-2017). https://seer.cancer.gov/data-software/documentation/seerstat/nov2019, 2019. Accessed: 2021-1-26.

[116] T. Therneau, C. Crowson, and E. Atkinson. Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model. https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf, 2020. Accessed: 2021-01-26.

[117] T. M. Therneau. *A Package for Survival Analysis in R*, 2020. R package version 3.2-7.

[118] T. M. Therneau and P. M. Grambsch. *Modeling survival data: Extending the Cox model*. Springer, 2000.

[119] T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[120] I. Thior, S. Lockman, L. M. Smeaton, R. L. Shapiro, C. Wester, S. J. Heymann, P. B. Gilbert, L. Stevens, T. Peter, S. Kim, et al. Breastfeeding plus infant zidovudine prophylaxis for 6 months vs formula feeding plus infant zidovudine for 1 month to reduce mother-to-child HIV transmission in Botswana. *JAMA*, 296(7):794–805, 2006.

[121] G. Tutz and H. Binder. Flexible modelling of discrete failure time including time-varying smooth effects. *Statistics in Medicine*, 23(15):2445–2461, 2004.

[122] P. J. Verweij and H. C. Van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314, 1993.

[123] P. J. M. Verweij and H. C. van Houwelingen. Time-dependent effects of fixed covariates in Cox regression. *Biometrics*, 51(4):1550–1556, 1995.

[124] M. Wang, L. Kong, Z. Li, and L. Zhang. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Statistics in Medicine*, 35(10):1706–1721, 2016.

[125] Y. Wang, B. Nan, and J. D. Kalbfleisch. Kernel estimation of bivariate time-varying coefficient model for longitudinal data with terminal event. https://doi.org/10.48550/arXiv.2111.04938, 2021. Accessed: 2022-02-26.

[126] T. Welchowski and S. Matthias. discSurv: Discrete time survival analysis. https://cran.R-project.org/package=discSurv, 2019. R package version 1.4.1.

[127] J. B. Wish. The role of 30-day readmission as a measure of quality. *Clinical Journal of the American Society of Nephrology*, 9(3):440–442, 2014.

[128] R. A. Wolfe, V. B. Ashby, E. L. Milford, A. O. Ojo, R. E. Ettenger, L. Y. Agodoa, P. J. Held, and F. K. Port. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England Journal of Medicine*, 341(23):1725–1730, 1999.

[129] W. H. Wong. Theory of partial likelihood. *Annals of Statistics*, 14(1):88–123, 1986.

[130] S. N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):413–428, 2000.

[131] S. N. Wood. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036, 2006.

[132] S. N. Wood. *Generalized Additive Models: An Introduction with R*. CRC Press, 2017.

[133] M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 95(7), 2017.

[134] W. Wu, G. W. Gremel, K. He, J. M. Messana, A. Sen, J. H. Segal, C. Dahlerus, R. A. Hirth, J. Kang, K. Wisniewski, T. Nahra, R. Padilla, L. Tong, H. Gu, X. Wang, M. Slowey, A. Eckard, X. Ding, L. Borowicz, J. Du, B. Frye, and J. D. Kalbfleisch. The impact of COVID-19 on post-discharge outcomes for dialysis patients in the United States: Evidence from Medicare claims data. *Kidney360*, 3(6):1047–1056, 2022.

[135] W. Wu, G. W. Gremel, K. He, J. M. Messana, A. Sen, J. H. Segal, C. Dahlerus, R. A. Hirth, J. Kang, K. Wisniewski, T. Nahra, R. Padilla, L. Tong, H. Gu, X. Wang, M. Slowey, A. Eckard, X. Ding, L. Borowicz, J. Du, B. Frye, and J. D. Kalbfleisch. The impact of COVID-19 on postdischarge outcomes for dialysis patients in the United States: Evidence from Medicare claims data. *Kidney360*, 3(6):1047–1056, 2022.

[136] W. Wu, J. M. Taylor, A. F. Brouwer, L. Luo, J. Kang, H. Jiang, and K. He. Scalable proximal methods for cause-specific hazard modeling with time-varying coefficients. *Lifetime Data Analysis*, 28(2):194–218, 2022.

[137] W. Wu, Y. Yang, J. Kang, and K. He. Improving large-scale estimation and inference for profiling health care providers. *Statistics in Medicine*, 41(15):2840–2853, 2022.

[138] L. Xia, K. He, Y. Li, and J. D. Kalbfleisch. Accounting for total variation and robustness in profiling health care providers. *Biostatistics*, kxaa024:1–17, 2020.

[139] J. Xu, J. D. Kalbfleisch, and B. Tai. Statistical analysis of illness–death processes and semi-competing risks data. *Biometrics*, 66(3):716–725, 2010.

[140] J. Yan and J. Huang. Model selection for Cox models with time-varying coefficients. *Biometrics*, 68(2):419–428, 2012.

[141] Y. Yang. *Novel Methods for Estimation and Inference in Varying Coefficient Models*. PhD thesis, University of Michigan, ProQuest LLC, 789 East Eisenhower Parkway, P.O. Box 1346, Ann Arbor, MI 48106–1346, 2020. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/163251/yuanyang_1.pdf?sequence=1.

[142] Z. Yu, L. Liu, D. M. Bravata, and L. S. Williams. Joint model of recurrent events and a terminal event with time-varying coefficients. *Biometrical Journal*, 56(2):183–197, 2014.

[143] X. Zhao, J. Zhou, and L. Sun. Semiparametric transformation models with time-varying coefficients for recurrent and terminal events. *Biometrics*, 67(2):404–414, 2011.

[144] J. Zhou, N.-Y. Wang, and N. Wang. Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23(1):25, 2013.

[145] H. Zhu, J. Fan, and L. Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098, 2014.

[146] H. Zhu, R. Li, and L. Kong. Multivariate varying coefficient model for functional responses. *Annals of Statistics*, 40(5):2634, 2012.

[147] D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Annals of Statistics*, 18(1):329–353, 1990.