Modeling Time-varying Effects with Large-scale Survival Data: An Efficient Quasi-Newton Approach

Kevin He
Department of Biostatistics, School of Public Health, University of Michigan
and
Yuan Yang
Department of Biostatistics, School of Public Health, University of Michigan
and
Yanming Li
Department of Biostatistics, School of Public Health, University of Michigan
and
Ji Zhu
Department of Statistics, University of Michigan
and
Yi Li
Department of Biostatistics, School of Public Health, University of Michigan

August 31, 2016

Abstract

Nonproportional hazards models often arise in biomedical studies, as evidenced by a recent national kidney transplant study. During the follow up, the effects of baseline risk factors, such as patients' commorbidity conditions collected at transplantation, may vary over time. To model such dynamic changes of covariate effects, time-varying survival models have emerged as powerful tools. However, traditional methods of fitting time-varying effects survival model rely on an expansion of the original dataset in a repeated measurement format, which, even with a moderate sample size, leads to an extremely large working dataset. Consequently, the computational burden increases quickly as the sample size grows, and analyses of a large dataset such as our motivating example defy any existing statistical methods and software. We propose a novel application of quasi-Newton iteration method to model time-varying effects in survival analysis. We show that the algorithm converges superlinearly and is

computationally efficient for large-scale datasets. We apply the proposed methods, via a stratified procedure, to analyze the national kidney transplant data and study the impact of potential risk factors on post-transplant survival.

Keywords: Quasi-Newton; Survival analysis; Spline-based methods; Time-varying effects.

1 Introduction

With the advent of the big data era, there is an emergence in developing computationally feasible methods. For instance, in large-scale time-to-event data, datasets are often extremely large in the number of observations. Moreover, the effects of baseline risk factors may vary during the follow-up period, resulting in a weakening or strengthening of associations over time. To model the dynamic changes of covariate effects, time-varying survival models have emerged as powerful tools. However, the difficulty of model construction increases drastically as the sample size grows, prohibiting their use in analyzing large-scale time-to-event data.

This paper is motivated by the analysis of the national kidney transplant data, supported in part by the Health Resources and Services Administration. Renal failure is one of the most common and severe diseases in the nation. In 2010, a total of 116,946 new cases were reported. Kidney transplantation is a preferred treatment for renal failure patients. To optimize the survival benefit of transplantation, an accurate post-transplant survival model is needed, for which Cox proportional hazards regression (Cox (1972)) has been widely employed. This proportional hazards model stipulates that the covariates used in the regression model have a multiplicative effect on the death hazard throughout the follow-up period. However, such a proportionality assumption is not often practical and may lead to misleading results. For instance, obesity, generally viewed as a risk factor for mortality, presents rather dynamic impact on survival (Dekker et al. (2008); Kalantar-Zadeh (2005); de Mutsert et al. (2007)): it may have a protective effect in the short run, but can be a risk factor after a long term of exposure. Ignoring the complex time-varying nature of covariate effects may obscure more complex associations between risk factors and outcomes. This problem is fundamentally important for evaluating post-transplant mortality, as the validity of the fitted model hinges upon correct model structures.

In the framework of survival analysis with time-varying effects, Zucker and Karr (1990) studied the mathematical properties of a penalized partial likelihood approach and showed that the solution is a cubic spline with knots at the unique failure times. Gray (1992, 1994) proposed using spline functions to model time-varying effects. Hastie and Tibshirani (1993) discussed models with varying coefficients. Verweij and van Houwelingen (1995) suggested

fitting time varying effects using a penalty on the likelihood to control the pattern of the time effect. Berger et al. (2003) proposed time varying effects models with the use of fractional polynomials as time functions.

These methods are applicable in studies with relatively small sample sizes, but present either computational or methodological limitations for large-scale survival analysis. Most existing methods rely on expanding the original dataset in a repeated measurement format, such as in a counting-process style. The time is divided into small time intervals where a single event occurs, and for each time interval, the covariate values and outcome in the interval for each subject still under observation are stacked to form a large working dataset. Even with a moderate sample size, such an expansion leads to an extremely large working dataset that overwhelms current computational capacity. For instance, a dataset with 5,000 events (assuming no ties) would lead to an expanded dataset with more than 12 million records, which would easily overwhelm a computer with an 8G memory.

An alternative approach based on the Kronecker product was suggested by Perperoglou et al. (2006) to avoid the expansion of a large-scale dataset. When the sample size is small (e.g., less than 10,000), the Kronecker product based Newton-Raphson algorithm, which involves computation of the Hessian matrices at each iterative step, can be applied to optimize the partial likelihood function. However, when the sample size is large (as in our motivating example), the computation of the Hessian matrix becomes very computationally expensive. When the number of parameters is also large, the inversion of the Hessian matrix may be impractical. The iterative computation and inversion of Hessian matrices can be very cumbersome in the Newton steps.

The remainder of this article is organized as follows. In Section 2, we first summarize notations and some requisite preliminaries. We then propose a new modeling approach based on a quasi-Newton method. The proposed algorithm improves upon the traditional methods by avoiding iterative computation and inversion of Hessian matrices. The approach is broadly applicable to large-scale time-to-event data with time-varying effects, for which no effective methods are currently available. Finally, we extend the proposed procedure to incorporate a multicenter data structure. Convergence properties of the proposed approaches are considered in Section 3. Finite-sample properties are examined in Section

4 through simulations. The proposed methods are applied to analyze the national kidney transplant data in Section 5, and we conclude the article with a discussion in Section 6.

2 Method

Let T_i and C_i represent the survival and censoring times, respectively, for the ith patient. The total number of subjects is denoted by n. Observation times are denoted by $X_i = T_i \wedge C_i$, where $a \wedge b = \min\{a, b\}$ and $I(A)$ is an indicator function taking the value 1 when condition A holds and 0 otherwise. The observed death indicators are denoted by $\Delta_i = I(T_i \leq C_i)$. Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iP})^T$ be a P-dimensional covariate vector. Let $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_P(t))^T$ be a $P-$dimensional vector of potentially time-varying coefficients. The observed data consist of n independent vectors, $(X_i, \Delta_i, \mathbf{Z}_i)$.

Let $\lambda(t|\mathbf{Z}_i)$ be the hazard function given $\mathbf{Z}_i$ and the time-varying effects survival model is stipulated as

$$\lambda(t|\mathbf{Z}_i) = \lambda_0(t) \exp(\mathbf{Z}_i^T \boldsymbol{\beta}(t)),$$

where $\lambda_0(t)$ is the baseline hazard. The corresponding log-partial likelihood (under noninformative censoring)

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left[\mathbf{Z}_i^T \boldsymbol{\beta}(T_i) - \log \left\{ \sum_{\ell \in R_i} \exp\left(\mathbf{Z}_\ell^T \boldsymbol{\beta}(T_i) \right) \right\} \right],$$

where $R_i = \{\ell : T_\ell \geq T_i\}$ is the at-risk set.

2.1 Estimation with B-spline

To estimate $\boldsymbol{\beta}$, a commonly applied approximation is to span $\boldsymbol{\beta}(\cdot)$ by a set of B-splines on a fixed grid of knots, usually taken to have an equal number of events within each interval Gray (1992). More specifically, for $p = 1, \ldots, P$, $\beta_p(\cdot)$ is an expansion of the form

$$\beta_p(t) = \boldsymbol{\theta}_p^T \mathbf{B}(t) = \sum_{k=1}^{K} \theta_{pk} B_k(t), \quad p = 1, \cdots, P,$$

where $\mathbf{B}(t) = (B_1(t), \ldots, B_K(t))^T$ forms a basis for a finite-dimensional space, and $\boldsymbol{\theta}_p = (\theta_{p1}, \ldots, \theta_{pK})^T$ is a vector of coefficients with θ_{pk} being the corresponding coefficient vector

for the kth component of the pth covariate. Consider a length-PK parameter vector $\boldsymbol{\theta} = vech(\boldsymbol{\Theta})$, the vectorization of $P \times K$-dimensional coefficient matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P)^T$ by row, which is a reparamerization of the original $\boldsymbol{\beta}$ on a space spanned by a set of B-splines. With that, the log-partial likelihood function is

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Delta_i \left[\mathbf{Z}_i^T \boldsymbol{\Theta} \mathbf{B}(T_i) - \log \left\{ \sum_{\ell \in R_i} \exp\left(\mathbf{Z}_\ell^T \boldsymbol{\Theta} \mathbf{B}(T_i) \right) \right\} \right]. \tag{1}$$

Corresponding to (1), the gradient is

$$\nabla l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Delta_i \left\{ \mathbf{Z}_i - \overline{\mathbf{Z}}_i(\boldsymbol{\Theta}, T_i) \right\} \otimes \mathbf{B}(T_i),$$

where $\otimes$ is the Kronecker product and

$$\overline{\mathbf{Z}}_i(\boldsymbol{\Theta}, T_i) = \frac{\sum_{\ell \in R_i} \mathbf{Z}_\ell \exp\{\mathbf{Z}_\ell^T \boldsymbol{\Theta} \mathbf{B}(T_i)\}}{\sum_{\ell \in R_i} \exp\{\mathbf{Z}_\ell^T \boldsymbol{\Theta} \mathbf{B}(T_i)\}}.$$

The Hessian matrix is

$$\nabla^2 l_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \Delta_i \mathbf{V}_i(\boldsymbol{\Theta}, T_i) \otimes \left\{ \mathbf{B}(T_i) \mathbf{B}^T(T_i) \right\}, \tag{2}$$

where

$$\mathbf{V}_i(\boldsymbol{\Theta}, T_i) = \frac{\mathbf{S}_i^{(2)}(\boldsymbol{\Theta}, T_i) S_i^{(0)}(\boldsymbol{\Theta}, T_i) - \{\mathbf{S}_i^{(1)}(\boldsymbol{\Theta}, T_i)\}^{\otimes 2}}{\{S_i^{(0)}(\boldsymbol{\Theta}, T_i)\}^2},$$

and

$$\mathbf{S}_i^{(r)}(\boldsymbol{\Theta}, T_i) = \sum_{\ell \in R_i} \exp\{\mathbf{Z}_\ell^T \boldsymbol{\Theta} \mathbf{B}(T_i)\} \mathbf{Z}_\ell^{\otimes r},$$

for $r = 0, 1, 2$. For a column vector $\mathbf{v}$, $\mathbf{v}^{\otimes 2}$ denotes the outer product $\mathbf{v}\mathbf{v}^T$.

It is worth noting that in linear or generalized linear models, estimation of time-varying effects can be achieved by including interactions between $\mathbf{Z}_i$ and $\mathbf{B}(T_i)$. In survival analysis, however, it is more challenge due to the cross-terms $\mathbf{Z}_\ell$ and $\mathbf{B}(T_i)$ (for all i and ℓ such that $\ell \in R_i$). The iterative computation and inversion of Hessian matrices can be very computationally demanding in the Newton steps. Numerically, while the computation of a large Hessian matrix is challenging, the gradients however are often much easier to compute. Therefore, a low-rank update approach is viable as it avoids computing the Hessian matrix

iteratively. The Hessian matrix is approximated by adding low-rank updates based on gradients, as the gradients provide information on the second derivative of the log-partial likelihood along the search direction. This motivates us to consider an approach based on the quasi-Newton method.

2.2 Review for Low-rank Update Methods to Approximate Hessian Matrices

Consider a first-order Taylor approximation of the log-partial likelihood at the mth iteration

$$\nabla l_n(\widehat{\boldsymbol{\theta}}^{(m+1)}) - \nabla l_n(\widehat{\boldsymbol{\theta}}^{(m)}) \approx \nabla^2 l_n(\widehat{\boldsymbol{\theta}}^{(m)})(\widehat{\boldsymbol{\theta}}^{(m+1)} - \widehat{\boldsymbol{\theta}}^{(m)}).$$

An update of the approximation to the Hessian matrix is required to be symmetric and satisfy the secant condition:

$$\mathbf{H}^{(m+1)} d^{(m)} = g^{(m)},$$

where $\mathbf{H}^{(m+1)}$ is the new Hessian approximation,

$$d^{(m)} = \widehat{\boldsymbol{\theta}}^{(m+1)} - \widehat{\boldsymbol{\theta}}^{(m)},$$

$$g^{(m)} = \nabla l_n(\widehat{\boldsymbol{\theta}}^{(m+1)}) - \nabla l_n(\widehat{\boldsymbol{\theta}}^{(m)}).$$

Furthermore, the difference between successive approximation $\mathbf{H}^{(m)}$ and $\mathbf{H}^{(m+1)}$ is assumed to be of low rank. Two of the most popular quasi-Newton methods are the rank-one updates (Davidon (1959)) and rank-two updates. Because the approximation of the Hessian matrix is not always negative definite with the rank-one updates, we focus on the rank-two updates, which was proposed by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970), known as the BFGS algorithm:

$$\mathbf{H}^{(m+1)} = \mathbf{H}^{(m)} + b^{(m)} g^{(m)} (g^{(m)})^T + c^{(m)} \mathbf{H}^{(m)} d^{(m)} (d^{(m)})^T \mathbf{H}^{(m)},$$

where $b^{(m)}$ and $c^{(m)}$ are chosen to satisfy the secant condition

$$b^{(m)} = 1/\{(g^{(m)})^T d^{(m)}\},$$

$$c^{(m)} = -1/\{(d^{(m)})^T \mathbf{H}^{(m)} d^{(m)}\}.$$

The BFGS updates generate a negative definite approximation whenever the initial approximation, $\mathbf{H}^{(1)}$, is negative definite and the curvature condition (Nocedal and Wright (2006)) is satisfied:

$$(g^{(m)})^T d^{(m)} < 0. \tag{3}$$

2.3 Quasi-Newton Estimation for Time-varying Effects

One important question in applying the quasi-Newton method is how to choose the initial approximation $\mathbf{H}^{(1)}$: an oversimplified matrix is convenient but may be poorly scaled and result in a slow convergence rate.

For estimating time-varying effects in survival analysis, we propose to estimate $\mathbf{H}^{(1)}$ as follows. Initialize $\widehat{\boldsymbol{\theta}}^{(0)} = \widetilde{\boldsymbol{\beta}} \otimes \mathbf{1}_K$, (e.g., $\hat{\theta}^{(0)}_{pk} = \tilde{\beta}_p$, for $k = 1, \ldots, K$ and $p = 1, \ldots, P$), where $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_P)$ is a vector of coefficients fitted from the Cox proportional hazards model and $\mathbf{1}_K$ is a $K \times 1$ vector with all element 1. Based on the property of B-splines,

$$\hat{\beta}^{(0)}_p(t) = \sum_k \tilde{\beta}_p B_k(t) = \tilde{\beta}_p.$$

Therefore, $\mathbf{Z}_\ell^T \widehat{\boldsymbol{\Theta}}^{(0)} \mathbf{B}(T_i)$ reduces to $\mathbf{Z}_\ell^T \widetilde{\boldsymbol{\beta}}$, and $\mathbf{V}_i(\widehat{\boldsymbol{\Theta}}^{(0)}, T_i)$ reduces to $\mathbf{V}_i(\widetilde{\boldsymbol{\beta}})$, where

$$\mathbf{V}_i(\widetilde{\boldsymbol{\beta}}) = \frac{\mathbf{S}_i^{(2)}(\widetilde{\boldsymbol{\beta}}) S_i^{(0)}(\widetilde{\boldsymbol{\beta}}) - \{\mathbf{S}_i^{(1)}(\widetilde{\boldsymbol{\beta}})\}^{\otimes 2}}{\{S_i^{(0)}(\widetilde{\boldsymbol{\beta}})\}^2}$$

is the corresponding quantity from the Cox proportional hazards model, with

$$\mathbf{S}_i^{(r)}(\widetilde{\boldsymbol{\beta}}) = \sum_{\ell \in R_i} \exp(\mathbf{Z}_\ell^T \widetilde{\boldsymbol{\beta}}) \mathbf{Z}_\ell^{\otimes r},$$

for $r = 0, 1, 2$. The initial estimation for the Hessian matrix approximation can be given by

$$\mathbf{H}^{(1)} \equiv \nabla^2 l_n(\widehat{\boldsymbol{\theta}}^{(0)}) = -\sum_{i=1}^n \Delta_i \mathbf{V}_i(\widetilde{\boldsymbol{\beta}}) \otimes \{\mathbf{B}(T_i)\mathbf{B}^T(T_i)\},$$

which avoids the computation of cross-terms between $\mathbf{Z}_\ell$ and $\mathbf{B}(T_i)$ for $\ell \in R_i$. Comparing to (2), the computational effort for $\mathbf{H}^{(1)}$ is light.

Moreover, to avoid inverting a large Hessian matrix, we utilize the quasi-Newton method which generates an approximation to the inverse of a Hessian matrix directly. With an initial Hessian matrix, $\mathbf{H}^{(1)}$, we use $\mathbf{K}^{(1)} = \{\mathbf{H}^{(1)}\}^{-1}$ to be the initial estimation of the inverse of the Hessian matrix. Denote the approximation to the inverse of a Hessian matrix at the current iteration by $\mathbf{K}^{(m)}$, the updated approximation of $\mathbf{K}^{(m+1)}$ can be given by

$$\mathbf{K}^{(m+1)} = \{\mathbf{I} - b^{(m)} g^{(m)} (d^{(m)})^T\}^T \mathbf{K}^{(m)} \{\mathbf{I} - b^{(m)} g^{(m)} (d^{(m)})^T\} + b^{(m)} d^{(m)} (d^{(m)})^T. \tag{4}$$

Based on the BFGS updates (4), the quasi-Newton iteration is then given by

$$\widehat{\boldsymbol{\theta}}^{(m+1)} = \widehat{\boldsymbol{\theta}}^{(m)} - \alpha^{(m)} (\mathbf{K}^{(m)}) \nabla ln(\widehat{\boldsymbol{\theta}}^{(m)}), \tag{5}$$

where $\alpha^{(m)}$ is positive (i.e., the step size). We implement a line search procedure such that the step size satisfies the strong Wolfe conditions (Dennis and Morè (1977))

$$l_n(\widehat{\boldsymbol{\theta}}^{(m+1)}) \geq l_n(\widehat{\boldsymbol{\theta}}^{(m)}) - c_1 \alpha^{(m)} (\mathbf{K}^{(m)}) \nabla l_n(\widehat{\boldsymbol{\theta}}^{(m)}), \tag{6}$$

$$\left| \nabla L(\widehat{\boldsymbol{\theta}}^{(m+1)})^T (\mathbf{K}^{(m)}) \nabla L(\widehat{\boldsymbol{\theta}}^{(m)}) \right| \leq c_2 \alpha^{(m)} \left| \nabla L(\widehat{\boldsymbol{\theta}}^{(m)}) (\mathbf{K}^{(m)}) \nabla ln(\widehat{\boldsymbol{\theta}}^{(m)}) \right|, \tag{7}$$

with $0 < c_1 < c_2 < 1$. As suggested by Lange (2012), typical values of c_2 are 0.9 and c_1 is chosen to be quite small, say 0.001. Because the log-partial likelihood function is continuously differentiable and bounded from above, an application of Lemma 3.1 of Nocedal and Wright (2006) entails that there exist intervals of step lengths that satisfy the strong Wolfe condition. The first part of the strong Wolfe condition (6) provides a sufficient increment in the log-partial likelihood function. The second part of the strong Wolfe condition (7) ensures the curvature condition (3) is satisfied and hence the $\mathbf{K}^{(m+1)}$ is negative definite. The iteration in (5) continues until convergence of $\boldsymbol{\theta}$ or the relative change in the log-partial likelihood is less than a convergence threshold (e.g., $1.0e^{-9}$). The convergence property of the quasi-Newton iteration is provided in Section 3. Alternatively, a potentially less optimal line search strategy, backtracking (e.g., if the initial increment of the parameter estimation does not produce a sufficient increase in the partial likelihood, then reduce the increment by half, and so forth), can be applied to simplify the computation. Our numerical experiments indicate that the backtracking approach is more sensitive to the choice of initial $H^{(1)}$. In some cases, a simple choice of identity matrix may suffer from non-convergence. In contrast, with the proposed initial estimation of the Hessian matrix, convergence is often achieved within $10 - 20$ steps.

2.4 Testing for Time-varying Effects

An ideal model construction procedure for time-varying effects should distinguish a subset of variables in the model with time-independent coefficients and those with time-varying coefficients.

Schoenfeld (1982) introduced Schoenfeld residuals to check proportional hazard assumptions. Grambsch and Therneau (1994) proposed scaling and smoothing these residuals for testing and plots to reveal the functional form of time-varying effects. The resulting method is the default for checking the proportional hazards assumptions in the statistical package Survival in R. As noted in Grambsch and Therneau (1994), the scaled Schoenfeld residuals are one-step Newton estimators with time-independent coefficients fitted from the Cox proportional hazards model as initial values. When the magnitude of true time-varying effects is small (e.g., initial values are close to the optimal), such an one-step estimator may provide a sound approximation. However, in more general cases the scaled Schoenfeld residuals may result in misleading estimation (more details are provided in Section 4).

We propose a test for time-varying effects based on quasi-Newton estimators. To test $H_0 : \beta_p(t) = \beta_p$, we specify a matrix $\mathbf{C}$ such that $\mathbf{C}\boldsymbol{\theta}_p = 0$ corresponds to the contrast that $\theta_{p1} = \cdots = \theta_{pK}$. A Wald test can be constructed by $(\mathbf{C}\boldsymbol{\theta}_p)^T Var(\boldsymbol{\theta}_p)(\mathbf{C}\boldsymbol{\theta}_p)$, where $Var(\boldsymbol{\theta}_p)$ is obtained through the quasi-Newton methods described in previous sections. The statistics are approximately chi-square distributed with $K-1$ degree of freedom under the null hypothesis. Simulations in Section 4 show that such statistics provide a good evaluation for the time-varying effects and can be particularly useful for analyzing large-scale data.

Finally, once we distinguish variables with time-independent coefficients and time-varying coefficients, we can fit the final model with linear equality constraints. For instance, with the constraint $\mathbf{C}\boldsymbol{\theta} = 0$, the revised updates can be obtained by the projection of the unconstrained increment onto the null space of $\mathbf{C}$

$$\widehat{\boldsymbol{\theta}}^{(m+1)} = \widehat{\boldsymbol{\theta}}^{(m)} - \alpha^{(m)}\{\mathbf{K}^{(m)} - \mathbf{K}^{(m)}\mathbf{C}^T(\mathbf{C}\mathbf{K}^{(m)}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{K}^{(m)}\}\nabla ln(\widehat{\boldsymbol{\theta}}^{(m)}).$$

2.5 Stratified Time-varying Effects Model

Another important consideration in large-scale biomedical studies is that the observations are often from multiple medical providers with large-scale electronic health records. In our motivating setting with renal failure patients, data are frequently derived from multiple dialysis facilities or transplant centers. In these studies, there is often much variation of practice patterns across medical centers. Some of the variation reflects the differential practice preferences of physicians from a specific region or center. In the absence of adjustment for center effects, the estimation of covariate effects may produce a substantially biased parameter estimates due to uncontrolled confounding by centers (Pan (2002)). Kalbfleisch and Wolfe (2013) suggested to use stratified models with center-specific baseline hazards to avoid confounding between patient characteristics and center effects. In large-scale time-to-event data, one advantage of using a stratified model is that it greatly reduces the number of calculations across the partial likelihood contributions. This advantage is especially important for large-scale data exemplified in our study.

We now extend the proposed quasi-Newton algorithm to stratified models. To incorporate the multicenter data structure, we modify the notations as follows. Let J be the number of transplant centers. The total number of subjects is denoted by $n = \sum_{j=1}^{J} n_j$, where n_j is the number of subjects in center j. The observed data consist of n independent vectors, $(X_{ij}, \Delta_{ij}, \mathbf{Z}_{ij})$, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$.

Let $\lambda_j(t|\mathbf{Z}_{ij})$ be the center-specific hazard function for center j given $\mathbf{Z}_{ij}$ and the stratified time-varying effects survival model is stipulated as

$$\lambda_j(t|\mathbf{Z}_{ij}) = \lambda_{0j}(t) \exp(\mathbf{Z}_{ij}^T \boldsymbol{\beta}(t)),$$

where $\lambda_{0j}(t)$ is the center-specific baseline hazard. The corresponding stratified log-partial likelihood (under noninformative censoring)

$$l_n(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Delta_{ij} \left[\mathbf{Z}_{ij}^T \boldsymbol{\beta}(T_{ij}) - \log \left\{ \sum_{\ell \in R_{ij}} \exp\left(\mathbf{Z}_{\ell j}^T \boldsymbol{\beta}(T_{ij}) \right) \right\} \right],$$

where $R_{ij} = \{\ell : 1 \leq \ell \leq n_j,\ T_{\ell j} \geq T_{ij}\}$ is the center-specific at-risk set.

With the original $\boldsymbol{\beta}$ spanned by a set of B-splines, the stratified log-partial likelihood

function is

$$l_n(\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Delta_{ij} \left[\mathbf{Z}_{ij}^T \boldsymbol{\Theta} \mathbf{B}(T_{ij}) - \log \left\{ \sum_{\ell \in R_{ij}} \exp\left(\mathbf{Z}_{\ell j}^T \boldsymbol{\Theta} \mathbf{B}(T_{ij}) \right) \right\} \right].$$

The remaining algorithms are the same as those in previous subsections.

3 Convergence Properties

The convergence properties of the proposed quasi-Newton algorithms are summarized by the following proposition. We defer all the regularity conditions and the proof to the Appendix.

Proposition 1 *The sequence* $\widehat{\boldsymbol{\theta}}^{(m)}$ *generated by the proposed quasi-Newton algorithm possesses a limit, and that limit is the optimal point,* $\boldsymbol{\theta}^\star$*, which maximizes the log-partial likelihood (1). Moreover, the* $\widehat{\boldsymbol{\theta}}^{(m)}$ *converges to* $\boldsymbol{\theta}^\star$ *at a superlinear rate; i.e.,* $||\widehat{\boldsymbol{\theta}}^{(m+1)} - \boldsymbol{\theta}^\star|| \leq o(||\widehat{\boldsymbol{\theta}}^{(m)} - \boldsymbol{\theta}^\star||)$.

4 Simulation Study

4.1 Evaluation of Computation Speed

We first assess the speed of our algorithm through a simple simulation study. We considered the sample size within each center $n_j = 500$. The number of center J ranged from 2 to 200 as we increased the total sample size n from $1,000$ to $100,000$. Death times were generated from the center-specific exponential model, $\lambda_j(t|\mathbf{Z}_{ij}) = \lambda_{0j}(t) \exp(\mathbf{Z}_{ij}^T \boldsymbol{\beta}(t))$. Twenty covariates were generated from independent normal distributions. Censoring times were generated from uniform distributions on $[0, u]$, with u chosen to yield approximately $20 - 30\%$ censoring proportions. Ten ($K = 10$) basis functions were used and the knots were chosen to have an equal number of events within each interval.

Table 1 compares the computation time for a full-step Newton-Raphson method based on an extended-data method (termed Unstratified NR and Stratified NR for approaches without and with stratification on centers), the proposed quasi-Newton approach (termed

Unstratified quasi-NR and Stratified quasi-NR for versions without and with stratification on centers). Convergence criterion were chosen as the maximal absolute change of $\boldsymbol{\theta}$ is less than 10^{-9}. These timings were taken on an Dell laptop (model XPS 15) with quad-core 2.1-GHz Intel Core i7-3612QM processor and 8GB RAM.

n	J	Unstratified NR	Stratified NR	Unstratified quasi-NR	Stratified quasi-NR
1,000	2	7.05 Minutes	2.99 Minutes	0.65 Minutes	0.45 Minutes
5,000	10	Fail	11.96 Minutes	5.74 Minutes	0.86 Minutes
10,000	20	Fail	Fail	24.79 Minutes	1.56 Minutes
100,000	200	Fail	Fail	36.68 Hours	0.17 Hours

Table 1: Comparisons of computation time: 1 simulation loop; number of covariates P=20; n=sample size; J=number of centers; The Newton method is based on extended data approach, and is estimated using R package *Survival*; Fail means the computation exceeds the computer's max memory capacity

4.2 Estimation of Time-varying Effects

Setting 1 We first considered a simulation setting where the hazard functions were equal across centers, e.g., death times were generated from a exponential distribution with parameter $\lambda_j = 1$, for centers $j = 1, \ldots, 5$. The number of subjects within each center was $n_j = 200$. Censoring distribution was the same as those in Section 4.1. Three covariate were generated from independent standard normal distributions. We let β_1 be a time-varying effect $\beta_1(t) = 3\sin(3\pi t/4)$. All other covariate effects are time-independent: $\beta_2 = 1$ and $\beta_3 = -1$. Each data configuration was replicated 500 times.

Figure 1 compare the method based on one-step Newton-Raphson, unstratified and stratified quasi-Newton methods. The one-step Newton method results in biased estimators, especially for covariates with time-independent effects (β_2 and β_3). The bias increases as the magnitude of the time-varying effect increases. This is due to that the one-step Newton method uses constant coefficients (fitted from the Cox model) as initial values, which does not incorporate the time-varying effects. In contrast, quasi-Newton estimators

are much more accurate. The full-step Newton method provides very similar results as the quasi-Newton. Hence, they are omitted. More results for comparing the unstratified and stratified quasi-Newton methods are displayed in the first half of Table 2. We report average bias, average mean square error (MSE), average empirical coverage probabilities (CP), average computation times, average iterations until convergence, empirical power (average proportion that the test rejected the null hypothesis for variables with time-varying effects) and the empirical Type-I error (average proportion that the test rejected the null hypothesis for variables with time-independent effects) for all estimated coefficients over 500 replicates. The reported bias, CP and MSE are the average of pointwise estimates over simulated time points. For example, for each simulation replication, the bias is calculated as the average of pointwise estimates over simulated time points, e.g.,

$$\text{bias} = \frac{\sum_{p=1}^{P}\sum_{i=1}^{n}(\hat{\beta}_p(T_i) - \beta_p(T_i))}{Pn},$$

where $P = 3$, $n = 1,000$, $\hat{\beta}_p(T_i)$ and $\beta_p(T_i)$ are the estimated and true coefficients, respectively, for the pth variable at time T_i. Then the average bias reported in Table 2 is the average of bias over 500 replications.

In terms of the asymptotic approximation, both estimators are sufficiently well-behaved, in the sense that the empirical CP are generally consistent with the nominal value of 0.95. The bias and type I errors are slightly larger than expected. One potential explanation is that in the late stage, the at-risk set is small, causing wide confidence intervals (as shown in Figure 1). As one would expect, the MSE of stratified method is slightly larger than the unstratified version, which indicates that the stratified method may suffer from loss of power when the true hazard functions are center-independent.

Setting 2 We next considered a simulation setting in which the hazard functions differed by center. Specifically, death times were generated from center-specific Weibull model,

$$\lambda_j(t|\mathbf{Z}_{ij}) = \alpha_j\gamma_j t^{\gamma_j - 1}\exp(\mathbf{Z}_{ij}^T\boldsymbol{\beta}(t))$$

where values of $\alpha_j = (0.2, 0.5, 1, 1.5, 2)$ and $\gamma_j = (0.8, 0.9, 1, 1.1, 1.2)$ for $j = 1, \ldots, 5$. The covariate distribution was chosen to be independent normal with center-dependent mean. Other set ups were the same as those in setting 1. We compare the performance in Table

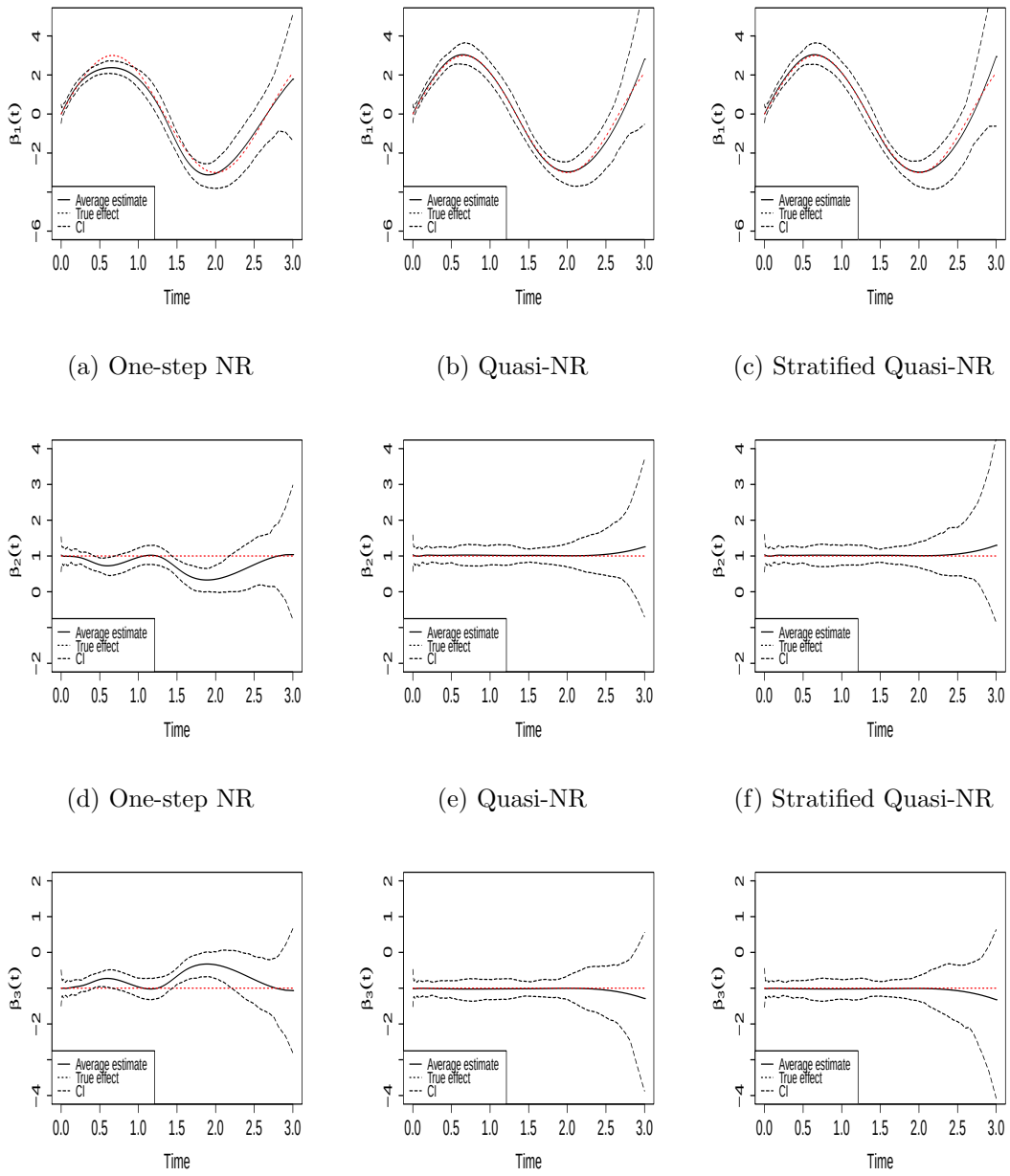
Setting	quasi-NR	Bias	MSE	CP	Time	Iteration	Power	Type-I Error
1	Unstratified	0.13	0.13	0.95	8.93 Seconds	18.51	1	0.08
	Stratified	0.12	0.14	0.95	4.80 Seconds	19.16	1	0.07
2	Unstratified	0.50	0.58	0.39	8.92 Seconds	16.82	1	0.56
	Stratified	0.11	0.06	0.94	5.48 Seconds	19.59	1	0.08

Table 2: Performance of unstratified and quasi-Newton methods; True effects: $\beta_1(t) = 3\sin(3\pi t/4)$, $\beta_2(t) = 1$ and $\beta_3(t) = -1$; sample size: n=1,000; number of centers: J=5 (200 subjects in each center); all centers have the same baseline hazard functions: exponential distribution with $\lambda = 1$; average bias: average of bias over simulated time points and across 500 replications; MSE: average mean square error; CP: average empirical coverage probabilities; Power: average proportion that the test rejected the null hypothesis for variables with time-varying effects; Type-I Error: average proportion that the test rejected the null hypothesis for variables with time-independent effects; 500 replications.

2 and Figure 3. The difference between the unstratified quasi-Newton and stratified quasi-Newton is quite pronounced. Since the unstratified model omits center effects, bias and MSE are quite large and the CP is notably less than 0.95. In contrast, stratified quasi-Newton is sufficiently accurate for this setting.

4.3 Testing for Time-varying Effects

Figure 3 shows the empirical power and the empirical Type-I error for tests based on proposed quasi-Newton and scaled Schoenfeld residuals (implemented by *cox.zph* in the R Survival package). The simulation setting up was similar to those in Setting 1, except for that $\beta_1(t) = \alpha\sin(3\pi t/4)$, where α varied between 0 and 3. The proposed testing outperforms the traditional method with higher power and smaller Type-I error (e.g., false positive). Detailed values for average P-values, empirical power and Type-I error can be found in Table A.3 of supplementary material. Figure A.1 in the supplementary material further investigated a setting with time-varying effects other than periodic pattern.

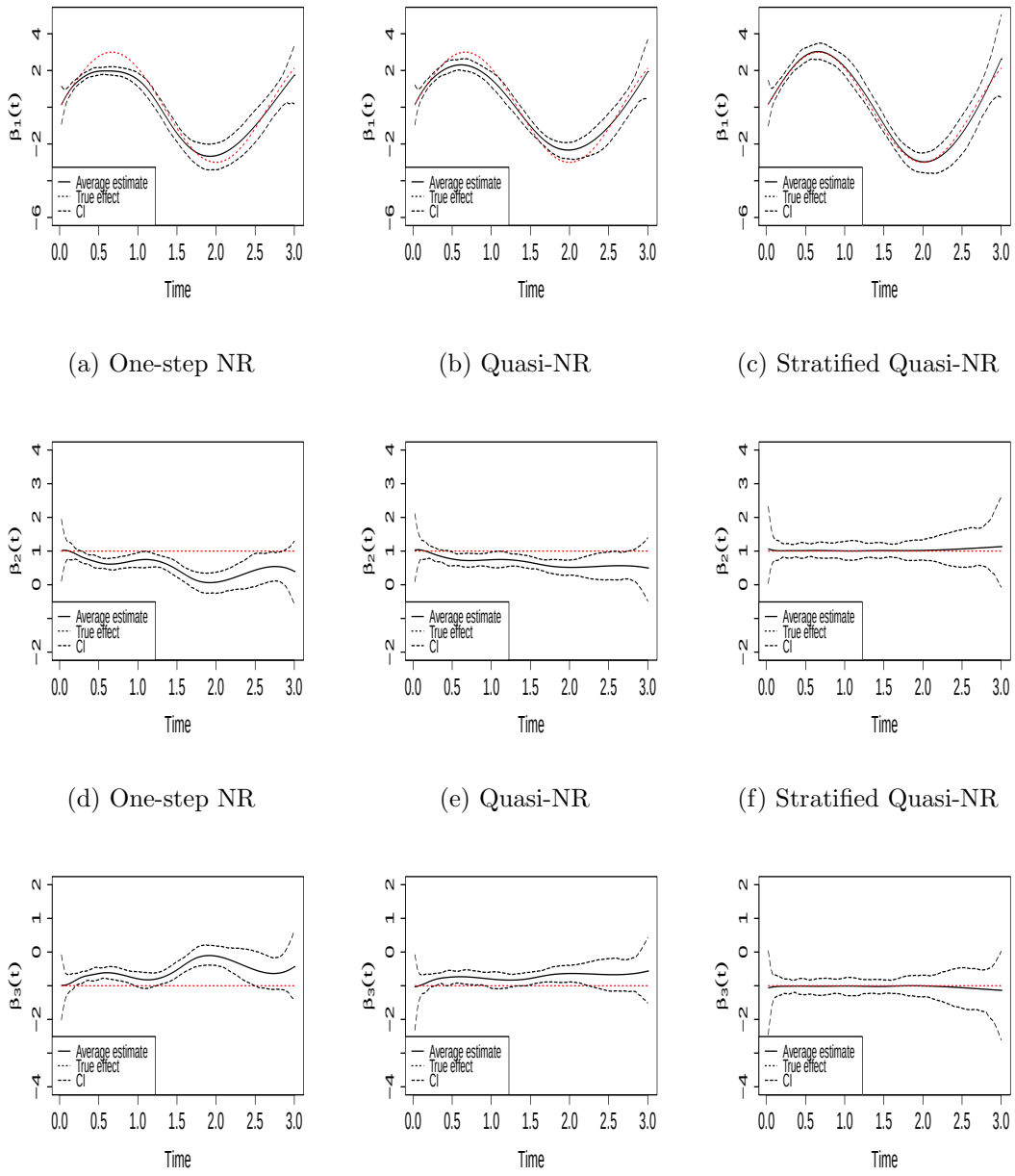


(a) One-step NR (b) Quasi-NR (c) Stratified Quasi-NR

(d) One-step NR (e) Quasi-NR (f) Stratified Quasi-NR

(g) One-step NR (h) Quasi-NR (i) Stratified Quasi-NR

Figure 1: Estimated coefficients for one-step Newton method (scaled Schoenfeld residuals), unstratified and stratified quasi-Newton methods. True effects: $\beta_1(t) = 3\sin(3\pi t/4)$, $\beta_2(t) = 1$ and $\beta_3(t) = -1$; sample size: n=1,000; number of centers: J=5 (200 subjects in each center); all centers have the same baseline hazard functions: exponential distribution with $\lambda = 1$; 500 replications.

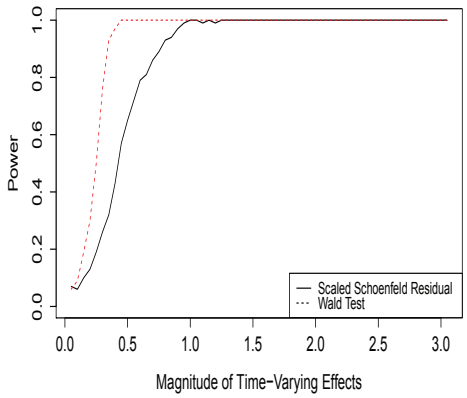


(a) One-step NR (b) Quasi-NR (c) Stratified Quasi-NR

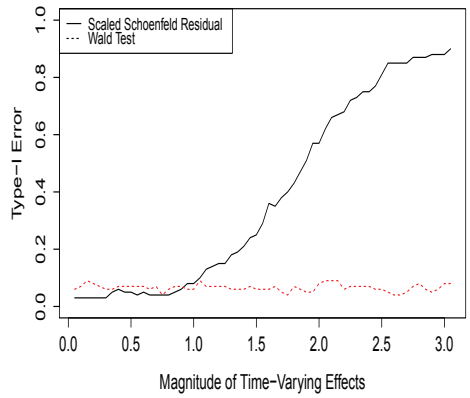
(d) One-step NR (e) Quasi-NR (f) Stratified Quasi-NR

(g) One-step NR (h) Quasi-NR (i) Stratified Quasi-NR

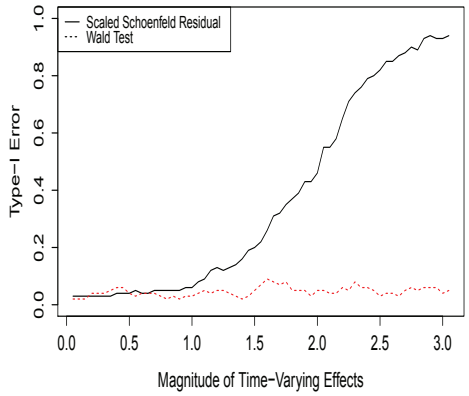
Figure 2: Estimated coefficients for one-step Newton method (scaled Schoenfeld residuals), unstratified quasi-Newton method and stratified quasi-Newton method. True effects: $\beta_1(t) = 3\sin(3\pi t/4)$, $\beta_2(t) = 1$ and $\beta_3(t) = -1$; sample size: n=1,000; number of centers: J=5 (200 subjects in each center); centers-specific baseline hazard functions follow Weibull distributions with parameters varying across centers; 500 replications.



(a) Power for β_1: time-varying effects



(b) Type-I Error for β_2: time-independent effects



(c) Type-I Error for β_3: time-independent effects

Figure 3: Power: average proportion that the test rejects the null hypothesis for variables with time-varying effects; Type-I error: average proportion that the test rejects the null hypothesis for variables with time-independent effects; True effects: $\beta_1(t) = \alpha \sin(3\pi t/4)$ where the magnitude of time-varying effects, α, varies from 0 to 3, $\beta_2(t) = 1$ and $\beta_3(t) = -1$; the tests for time-varying effect are described in Section 2.4; Type-I errors for quasi-Newton methods in Figures 3 (b) and (c) are around 0.05; detailed values for average P-values, empirical power and Type-I error can be found in Table A.3 of supplementary material.

5 Analysis

The motivating data were obtained from the Organ Procurement and Transplantation Network (OPTN), administered under a contract with the U.S. Department of Health and Human Services (HHS). Included in the analysis were adult renal failure patients ($\geq$ 18 years of age) who underwent deceased donor kidney transplantation between January 1994 and December 2012. Graft failure was considered to occur when the transplanted kidney ceased to function. Failure time (recorded in years) was defined as the time from transplantation to graft failure or death, whichever occurred first, with censoring at the end of study.

The final sample size is $n = 146,331$ from $J = 216$ centers. The median follow-up time is 11 years. The overall censoring rate is 62%. Adjustment covariates in this study included age, race, gender, donation after cardiac death (DCD), Expanded Criteria Donor (ECD), BMI (underweight, normal, overweight and obesity, where normal is the reference group), dialysis time, indicator of previous kidney transplant, cold ischemic time, cormorbidity conditions (e.g., glomerulonephritis, polycystic kidney disease, diabetes, hypertension).

Our analysis is to address two questions. First, we investigate whether the proportional hazards assumption is valid. Second, if the assumption is not valid, we estimate the appropriate functional form to explain the time-varying effect. As shown in Figure A.2, center-specific cumulative hazard functions varied across centers. The proposed stratified quasi-Newton method was implemented with 10 basis functions (knots were chosen to have an equal number of events within each interval). Figures 4-5 show a subset of fitted time-varying coefficients with 95% point-wise confidence intervals. The Wald test discussed in Section 2.4 was used to calculate p-values and identify factors that appear to have nonproportional hazards effects. As a comparison, a test based on scaled Schoenfeld residuals was carried out for each covariate. The p-values are listed in the caption. These results suggest that the effect of diabetes, male and black race (Figures 4a-4c) vary over time, resulting in strengthening associations over time. Conversely, Hispanic and ECD (Figures 4d-4e) have weakening association over time. Overweight and obesity (Figures 5b-5c) have protective effects in the short-term and then become risk factors after a long time exposure. Finally, the results for glomerulonephritis, polycystic kidney disease, and hypertension should be

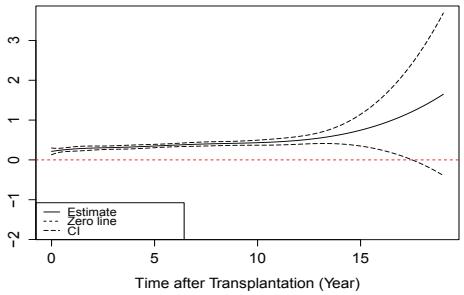
interpreted with cautions. Though the tests based on scaled Schoenfeld residuals indicate that the time-varying effects for these variables are significant, their time-varying effects turn out to be non-significant based on the quasi-Newton method. As shown in Figures 5d-5f, their effects are constant in the early stage of the follow-up period. In the late stage, although an increasing trend in the coefficients is observed, the at-risk set is small, causing wide confidence intervals.

6 Discussion

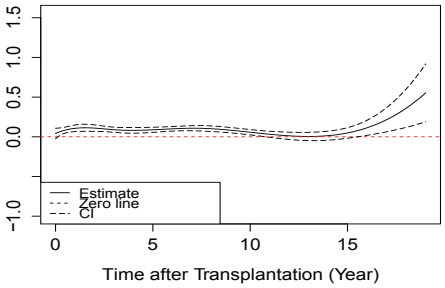
Statistical analysis of large-scale time-to-event data with potential time-varying effects presents daunting statistical challenges as well as exciting opportunities. Maximization of the partial likelihood via the traditional Newton method requires evaluation of the gradient and the Hessian matrix at each iteration. However, in large-scale analysis with time-varying effects, the numerical calculations and inversion of the Hessian matrix may have unreasonable costs, or may even be impractical. To improve computation efficiency, we propose a quasi-Newton approach that avoids iterative computation and inversion of the complex Hessian matrix. The proposed methods work efficiently for large-scale problems where current methods are impractical or even fail completely.

The proposed methods in this study are based on B-splines. A remaining question is how to choose the number and locations of knots for spline-based functions. Our simulations (Figures A.3 and A.4 in the Web_Supp.pdf file of Supplementary Material) confirm the advantage of previous recommendation by Gray (1992) to use splines with a moderate number of knots (e.g., 10). Moreover, our results (Figure A.5 in the Web_Supp.pdf file of Supplementary Material) show that the alternative approach, in which the knots are chosen to be equally spaced to cover the time-span, tends to be unstable in the right tail of the follow-up period. In contrast, the approach suggested by Gray (1992), for which the knots are chosen to include an equal number of events within each interval, offers a more stable estimation.

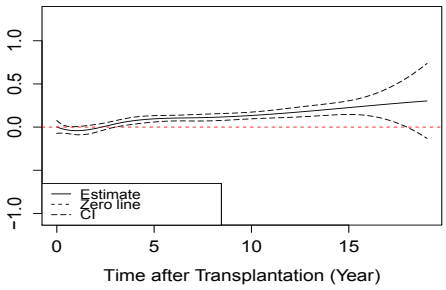
The stratified model is a useful tool in some applications, such as in analysis of large-scale electronic health records, where much variation of practice patterns often arises across regions or medical centers. In such cases, the stratified model ensures that adjustment



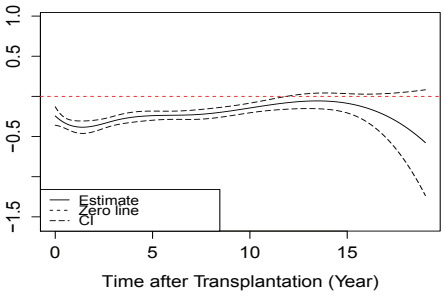
(a) Diabetes (p-value for test based on quasi-Newton < 0.001; p-value for test based on scaled Schoenfeld residuals < 0.001)



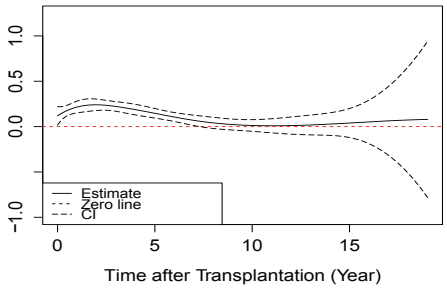
(b) Male (p-value for time-varying effect based on quasi-Newton $= 0.021$; p-value for test based on scaled Schoenfeld residuals < 0.001)



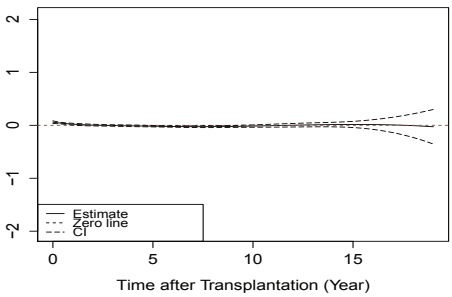
(c) Race: Black (p-value for test based on quasi-Newton < 0.001; p-value for test based on scaled Schoenfeld residuals < 0.001)



(d) Race: Hispanic (p-value for test based on quasi-Newton < 0.001; p-value for test based on scaled Schoenfeld residuals < 0.001)

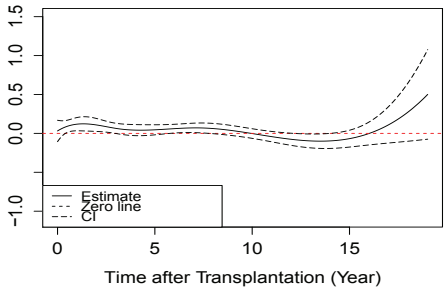


(e) Expanded Criterion Donor (p-value for time-varying effect based on quasi-Newton < 0.001; p-value for time-varying effect based on scaled Schoenfeld residuals < 0.001)

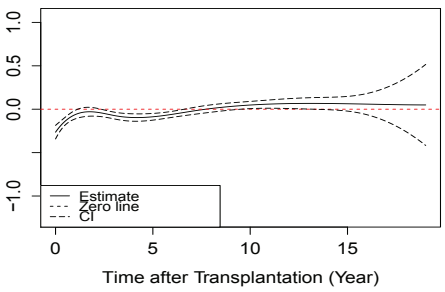


(f) Previous Kidney Transplant (p-value for test based on quasi-Newton $= 0.207$; p-value for test based on scaled Schoenfeld residuals $= 0.790$)

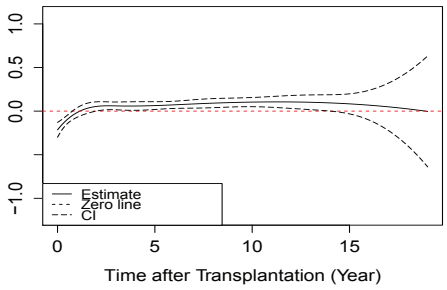
Figure 4: Real Data Application: Estimated coefficients are based on the quasi-Newton method (Section 2.3); 95% confidence interval (CI) is based on the delta method; The tests for time-varying effects are described in Section 2.4.



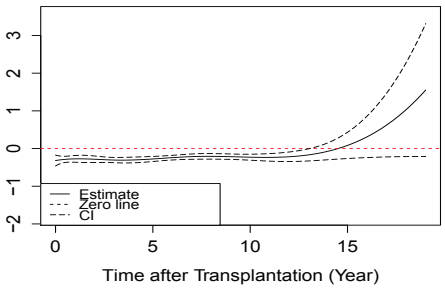
(a) Under-weight (BMI< 20) (p-value for test based on quasi-Newton = 0.033; p-value for test based on scaled Schoenfeld residuals = 0.022)



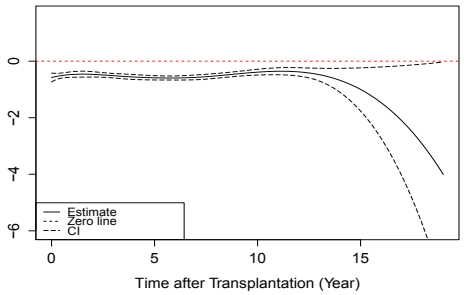
(b) Over-weight (25 <BMI< 30) (p-value for test based on quasi-Newton < 0.001; p-value for test based on scaled Schoenfeld residuals < 0.001)



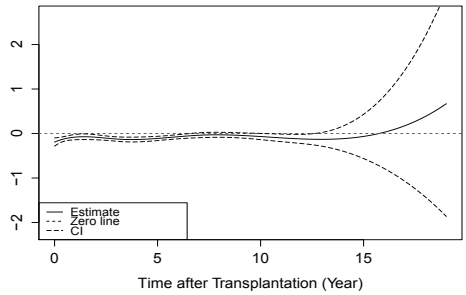
(c) Obesity (BMI> 30) (p-value for test based on quasi-Newton < 0.001; p-value for test based on scaled Schoenfeld residuals = 0.003)



(d) Glomerulonephritis (p-value for test based on quasi-Newton = 0.171; p-value for test based on scaled Schoenfeld residuals < 0.001)



(e) Polycystic Kidney Disease (p-value for test based on quasi-Newton = 0.059; p-value for test based on scaled Schoenfeld residuals < 0.001)



(f) Hypertension (p-value for test based on quasi-Newton = 0.103; p-value for test based on scaled Schoenfeld residuals < 0.001)

Figure 5: Real Data Application (continued)

covariate effects are not confounded by center effects. However, as pointed out by one reviewer, it implicitly introduces center-specific baseline hazards and hence a more complicated model, often leading to power loss compared with the unstratified model. Therefore, it should be applied with caution. In practice, center-specific baseline hazards can be assessed to determine whether stratifications are necessary. Finally, we remark that the proposed quasi-Newton method can be extended to incorporate time-dependent covariates. We will report this extension elsewhere.

Appendix

A1. Regularity Conditions:

To derive the convergence properties for the proposed algorithm, we impose the following regularity conditions. These conditions are commonly assumed in literatures and are applicable in most practical applications (Lin et al., 2000).

(a) $(X_i, \Delta_i, \mathbf{Z}_i)$ are independent and identically distributed random vectors.

(b) $P(X_i \geq \tau) > 0$ where τ is a pre-specified time point.

(c) The support of $\mathbf{Z}$ is bounded and $\boldsymbol{\theta}$ lies in a compact set.

(d) The Hessian matrix $\nabla^2 l_n(\boldsymbol{\theta}^\star)$ is negative definite at the stationary point $\boldsymbol{\theta}^\star$ such that $\nabla l_n(\boldsymbol{\theta}^\star) = 0$.

A2. Proof of Proposition 1:

First, conditions (b) and (c) lead to the boundedness of several quantities (e.g., log-partial likelihood, its gradient and the Hessian matrix). Given the fact that the log-partial likelihood function is concave and twice continuously differentiable, these boundedness properties and Condition (d) guarantee that l_n has a unique maximizer. In addition, continuity of the log-partial likelihood function and condition (c) implies that the super-level set $\{\boldsymbol{\theta} : l_n(\boldsymbol{\theta}) \geq l_n(\widehat{\boldsymbol{\theta}}^{(0)})\}$, for an arbitrary starting point $\widehat{\boldsymbol{\theta}}^{(0)}$, is a closed subset of a compact set and hence is compact. With the line search procedure described in Section 2.3, and the initial estimation for the inverted Hessian matrix approximation, $\mathbf{K}^{(1)}$, provided in

Section 2.3, all the consecutive $\mathbf{K}^{(m)}$ is negative definite and the quasi-Newton updates in (5) is an ascent direction. With Theorem 1 of Powell (1976), the sequence $\widehat{\boldsymbol{\theta}}^{(m)}$ generated by the proposed quasi-Newton algorithm possesses a limit, and that limit is the optimal point, $\boldsymbol{\theta}^{\star}$, which maximizes the log-partial likelihood. Therefore, the fixed point of the quasi-Newton iteration (i.e., $\widehat{\boldsymbol{\theta}}^{(m+1)} = \widehat{\boldsymbol{\theta}}^{(m)}$) is the stationary point of the log-partial likelihood function.

Finally, we show the convergence rate is superlinear. Continuity of the Hessian matrix and condition (d) imply that for all $\boldsymbol{\theta}$ sufficiently close to $\boldsymbol{\theta}^{\star}$, $\nabla^2 l_n(\boldsymbol{\theta})$ is negative definite. Moreover, convergence and ascent properties of quasi-Newton algorithm guarantee that there exist an iteration m' such that $\nabla^2 l_n(\boldsymbol{\theta})$ is negative definite for all $\boldsymbol{\theta} \in \Gamma = \{\boldsymbol{\theta} : l_n(\boldsymbol{\theta}) \geq l_n(\widehat{\boldsymbol{\theta}}^{(m')})\}$, and $\widehat{\boldsymbol{\theta}}^{(m)} \in \Gamma$ for all $m \geq m'$. An application of Weierstrass's theorem guarantees that there exist positive constants K_1 and K_2 such that

$$K_1 \mathbf{I} \preceq -\nabla^2 l_n(\boldsymbol{\theta}) \preceq K_2 \mathbf{I} \tag{8}$$

for all $\boldsymbol{\theta} \in \Gamma$, where matrices $\mathbf{A} \preceq \mathbf{B}$ denotes that $\mathbf{B} - \mathbf{A}$ is positive definite. The left inequality in (8) leads to the strong concavity of the log-partial likelihood.

Using the Taylor's theorem, we have that

$$\nabla l_n(\widehat{\boldsymbol{\theta}}^{(m+1)}) = \nabla l_n(\widehat{\boldsymbol{\theta}}^{(m)}) + \nabla^2 l_n(\widehat{\boldsymbol{\theta}}^{(m)})(\widehat{\boldsymbol{\theta}}^{(m+1)} - \widehat{\boldsymbol{\theta}}^{(m)}) + o(||\widehat{\boldsymbol{\theta}}^{(m+1)} - \widehat{\boldsymbol{\theta}}^{(m)}||).$$

The strong concavity property implies that there exist a positive constant K_3 such that

$$(g^{(m)})^T d^{(m)} = \{\nabla l_n(\widehat{\boldsymbol{\theta}}^{(m+1)}) - \nabla l_n(\widehat{\boldsymbol{\theta}}^{(m)})\}^T (\widehat{\boldsymbol{\theta}}^{(m+1)} - \widehat{\boldsymbol{\theta}}^{(m)}) \leq -K_3 ||\widehat{\boldsymbol{\theta}}^{(m+1)} - \widehat{\boldsymbol{\theta}}^{(m)}||^2 < 0.$$

Hence, for all $m \geq m'$, the curvature condition (3) is satisfied for quasi-Newton $\widehat{\boldsymbol{\theta}}^{(m)}$ with step size $\alpha^{(m)} = 1$. In addition, conditions (b)-(d) guarantee that the Hessian matrix $\nabla^2 l_n$ is Lipschitz continuous at the optimal $\boldsymbol{\theta}^{\star}$, i.e.,

$$||\nabla^2 l_n(\boldsymbol{\theta}) - \nabla^2 l_n(\boldsymbol{\theta}^{\star})|| \leq K_L ||\boldsymbol{\theta} - \boldsymbol{\theta}^{\star}||$$

for all $\boldsymbol{\theta}$ sufficiently close to $\boldsymbol{\theta}^{\star}$, where K_L is a positive constant. With Theorem 2 of Powell (1976), the sequence $\widehat{\boldsymbol{\theta}}^{(m)}$ converges to $\boldsymbol{\theta}^{\star}$ at a superlinear rate.

SUPPLEMENTARY MATERIAL

R codes used in Section 4 are contained in the zip file *QuasiNewton_1.0.zip* available online. An R package will soon be uploaded to the CRAN repository. Additional results referenced in Sections 4-6 can be found in the Web_Supp.pdf file. For reasons of confidentiality, the data analyzed in Section 5 is not publishable. The complete data set can be requested from the Organ Procurement and Transplantation Network (*https://optn.transplant.hrsa.gov/*).

ACKNOWLEDGEMENT

The authors are grateful to the Editor, the Associate Editor and three referees for their constructive comments and suggestions. The research of Yi Li was partially supported by a grant from the National Natural Science Foundation of China (No.11528102).

References

Berger, U., J. Schäer, and K. Ulm (2003). Dynamic cox modelling based on fractional polynomials: time-variations in gastric cancer prognosis. Statistics in Medicine 22(7), 1163–1180.

Broyden, C. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. Journal of Apply Mathematics 6(1), 76–90.

Cox, D. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187–200.

Davidon, W. (1959). Variable metric methods for minimization. In AEC Research and Development Report ANL–5990. Argonne National Laboratory, Argonne.

de Mutsert, R., M. Snijder, and van der Sman-de Beer F. et al. (2007). The 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, 967–974.

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

Dennis, J. and J. Morè (1977). Quasi-newton methods, motivation and theory. SIAM Review 19, 46–89.

Fletcher, R. (1970). A new approach to variable metric algorithms. Computer Journal 13(3), 317–322.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means. Mathematics of Computation 24(109), 23–26.

Grambsch, P. and T. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81, 515–526.

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

Gray, R. (1994). Spline-based tests in survival analysis. Biometrics 50(3), 640–652.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. Journal of the Royal Statistical Society, Series B 55, 757–796.

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

Kalbfleisch, J. and R. Wolfe (2013). On monitoring outcomes of medical providers. Statistics in Biosciences 5(2), 286–302.

Lange, K. (2012). In Optimization, 2nd Edtition. Springer Texts in Statistics.

Lin, D., L. Wei, I. Yang, and Z. Ying (2000). Semiparametic regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society, Series B 62, 711–730.

Nocedal, J. and S. Wright (2006). In Numerical optimization, 2nd edition. Springer, New York.

Pan, W. (2002). A note on the use of marginal likelihood and conditional likelihood in analyzing clustered data. The American Statistican 56, 171–174.

Perperoglou, A., S. le Cessie, and H. van Houwelingen (2006). A fast routine for fitting cox models with time varying effects of the covariates. Computer Methods and Programs in Biomedicine 25, 154–161.

Powell, M. (1976). Some global convergence properties of a variable metric algorithm without exact line searches. In Nonlinear Programming. SIAM-AMS Procedding 9.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika 69(1), 239–241.

Shanno, D. (1970). Conditioning of quasi-newton methods for function minimization. Math Comput 24(111), 647–656.

Verweij, P. and H. van Houwelingen (1995). Time-dependent effects of fixed covariates in cox regression. Biometrics 52, 1550–1556.

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