Efficient Assessment of Individualized Disease Risk and Treatment Response via Augmentation

Efficient Assessment of Individualized Disease Risk and Treatment Response via Augmentation

A dissertation presented

by

Yu Zheng

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2017

Dissertation Advisor: Professor Tianxi Cai Yu Zheng

Efficient Assessment of Individualized Disease Risk and Treatment Response via Augmentation

Abstract

T-year survival, defined as the survival status by a pre-specified time point t, is of great interest in many medical research areas. When the t-year survival is the outcome of interest in the individualized medicine, baseline covariates are used to predict the t-year survival for potential treatment response comparison. Time-specific generalized linear models estimated with inverse censoring probability weighting provides more robustness to model misspecification compared to other methods, but some challenges remain in the heavy censoring settings: the prediction model could be quite inefficient and deriving the optimal individualized treatment rules based on maximizing the population average survival probability could be difficult. Chapter 1 presents an imputation-based method to improve the efficiency of the baseline prediction model by incorporating the information from subjects censored before t and auxiliary covariates including the post-baseline secondary outcomes collected before censoring. Chapter 2 extends the method in Chapter 1 to incorporate the post-baseline intermediate covariates that are collected before t but have non-negligible missing values due to censoring. Chapter 3 proposes a systematic approach to derive optimal individualized treatment rules that maximizes the population average survival probability, and imputation-based augmentation approach is also developed to improve the efficiency of the estimation.

Contents

Acknowledgments

I want to express my special thanks to my dissertation advisor, Tianxi Cai, who guided me and encouraged me through this memorable journey of my life. I am also very grateful for Michael Hughes, who has been a great mentor through my entire career and Hajime Uno, an valuable committee member for my dissertation. This work is devoted to my dear parents Zongzhong Zheng and Xiaoqing Ouyang, my husband Yang Su, and my daughter Ivy Su. Their unconditional support enables me to fulfill my dream.

Chapter 1

Augmented Estimation for t-year Survival with Censored Regression Models

Abstract

Reliable and accurate risk prediction is fundamental for successful management of clinical conditions. Estimating comprehensive risk prediction models precisely, however, is a difficult task, especially when the outcome of interest is time to a rare event and the number of candidate predictors, p, is not very small. Another challenge in developing accurate risk models arises from potential model misspecification. Time-specific generalized linear models estimated with inverse censoring probability weighting are robust to model misspecification, but may be inefficient in the rare event setting. To improve the efficiency of such robust estimation procedures, various augmentation methods have been proposed in the literature. These procedures can also leverage auxiliary variables such as intermediate outcomes that are predictive of event risk. However, most existing methods do not perform well in the rare event setting, especially when p is not small. In this paper, we propose a two-step, imputation-based augmentation procedure that can improve estimation efficiency and that is robust to model misspecification. We also develop regularized augmentation procedures for settings where p is not small, along with procedures to improve the estimation of individualized treatment effect in risk reduction. Numerical studies suggest that our proposed methods substantially outperform existing methods in efficiency gains. The proposed methods are applied to an AIDS clinical trial for treating HIV-infected patients.

1.1 Introduction

Accurate and reliable risk prediction is fundamental for successful disease management, enabling those at high risk of early disease onset to be monitored more closely, and more aggressive treatment options to be considered for patients with poor prognoses. Precise estimation of risk prediction models in the survival setting, however, is difficult especially when the outcome is rare and the number of candidate predictors, p, is not very small. For example, long-term treatment-related co-morbidities, such as Tenofovir-associated renal disease in HIV-infected patients, have received increasing attention recently. When initiating a new therapy for HIV-infected patients, it would be of great value to accurately predict the likelihood of patients experiencing long-term treatment-related co-morbidities. However, obtaining a precise risk estimate for these co-morbidities based on clinical studies is challenging due to the limited number of observed events during follow up. Another challenge in developing risk models arises from potential model misspecification. For example, under the commonly used Cox proportional hazards (PH) model (Cox, 1972), regression coefficients can be interpreted as relative risk measures and can reflect the prognostic potential of predictors. With the unknown parameters in the model estimated, it is possible to predict subject-specific survival probabilities for risk classification. However, the PH assumption may not hold in many applications, and prediction models derived under model misspecification may not be validated in future studies since model estimates depend on censoring distributions (Cai and Cheng, 2008; Van Houwelingen, 2007), which are bound to differ across studies.

These challenges signify the importance of developing robust and efficient approaches to the estimation of risk prediction models. Here, we are specifically interested in predicting the risk of a clinical event occurring by time τ using baseline covariates through a τ-year time-specific generalized linear model (τ-GLM), as previously considered in Uno et al. (2007). In the presence of independent censoring and

possible model misspecification, Uno et al. (2007) showed that τ-GLM can be estimated consistently using inverse probability weighting (IPW). A risk prediction model is said to be estimated consistently if the model estimates converge in probability to the solution of a limiting estimating equation that is censoring-free. However, such a simple IPW estimator is not efficient as it only uses information from subjects who are not censored prior to τ. Incorporating information from censored subjects could potentially improve estimation efficiency. Additionally, auxiliary variables, including post-baseline intermediate outcomes that are predictive of the primary outcome, may be useful to further improve efficiency (Robins and Rotnitzky, 1992; Gray, 1994; Lu and Tsiatis, 2008; Parast et al., 2014).

There is a rich literature on improving estimation efficiency within an IPW framework by making use of information on auxiliary variables. However, most existing work focuses on non-censored data and simple population parameters such as the mean. Robins et al. (1994) propose a general approach for efficiency augmentation by estimating the selection probability as a function of covariates and extraneous surrogates. However, such a method suffers from the curse of dimensionality (Robins and Ritov, 1997), which would be amplified in the rare event setting. To improve robustness, augmented IPW (AIPW) approaches have been proposed for analyzing survival data (Scharfstein et al., 1999; Bang and Tsiatis, 2000; Tsiatis, 2006; Robins and Wang, 2000), including the τ-GLM (DiRienzo, 2009). The AIPW estimator, involving two separate models that estimate IPW weights and impute outcome, is doubly robust in the sense that it is asymptotically consistent and normally distributed as long as one of the two models is correctly specified. However, the efficiency gain from the AIPW method is often limited by misspecification of the imputation model (Robins et al., 2007). Furthermore, none of the existing methods work well when p is not small.

To overcome the challenges of model misspecification and the curse of dimensionality, we consider a two-step, imputation-based augmentation procedure to improve the estimation of the τ-GLM. In step I, we use a standard IPW approach to ap-

proximate the conditional risk function given baseline and surrogate variables under a flexible imputation model. In step II, we impute the τ-year event status based on the estimated conditional risk for each subject, and fit the regression model with imputed outcome using all subjects. Throughout, the risk function of interest is the τ-year risk given the baseline predictors, although post-baseline variables may be used for efficiency augmentation. We demonstrate that with properly chosen imputation procedures, our proposed augmented estimator is consistent under possible misspecification of both the τ-GLM and the imputation model. To further improve efficiency, we construct a combined estimator that is an optimal linear combination of the standard IPW estimator and the augmented estimator. Regularized augmentation procedures are also developed to allow p to remain moderate relative to the number of observed events.

The rest of the paper is organized as follows: in Section 2, we describe the proposed estimation procedure for both risk prediction and the extension to individualized treatment selection settings, where interest lies in estimating subject-specific treatment effects in risk reduction. In Section 3, we present results from simulation studies that demonstrate substantial potential gains in efficiency of estimation under a variety of settings. We illustrate our proposed procedures in Section 4 using an AIDS Clinical Trial for treating HIV patients. Some concluding remarks are given in Section 5.

1.2 Estimation

Let $T^\dagger$ be a continuous failure time, $\mathbf{X}$ be a $p \times 1$ vector of bounded baseline predictors, and C be the censoring variable. For convenience, we assume that $\mathbf{X} = (1, X_2, ..., X_p)^\intercal$. Let $\{(T_i^\dagger, \mathbf{X}_i, C_i), i = 1, ...,\text{n}\}$ be n independent copies of $(T^\dagger, \mathbf{X}, C)$. Due to censoring, for the i^{th} subject, we only observe $(T_i, \mathbf{X}_i, \delta_i)$, where $T_i = min(T_i^\dagger, C_i), \delta_i = I(T_i^\dagger \leq C_i)$ and $I(\cdot)$ is the indicator function. We focus on the setting where

$T_i^\dagger$ is subject to heavy censoring, but a $q \times 1$ vector post-baseline outcome, $\mathbf{S}$, is collected before τ and not subject to censoring. We first assume that C is independent of $\{T^\dagger, \mathbf{Z} = (\mathbf{X}^\intercal, \mathbf{S}^\intercal)^\intercal\}$, with a common survival distribution $G(\cdot)$. Extensions to incorporate covariate dependent censoring will be discussed in **??**. Let $\mathscr{D} = \{(T_i, \delta_i, \mathbf{Z}_i^\intercal)^\intercal, i = 1, ..., n\}$ denote the observed data. Throughout, for any vector $\boldsymbol{a} = (a_1, ..., a_p)^\intercal$, $\boldsymbol{a}_{[-1]} = (a_2, ..., a_p)^\intercal$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\intercal$.

1.2.1 Estimation procedure

To predict τ-year survival $Y_\tau = I(T^\dagger \leq \tau)$ based on $\mathbf{X}$, we fit a τ-GLM with logistic link:

$$P(T^\dagger \leq \tau|\mathbf{X}) = P(Y_\tau = 1|\mathbf{X}) = g(\boldsymbol{\beta}_\tau^\intercal \mathbf{X}), \quad \text{with } g(x) = e^x/(1+e^x), \tag{1.1}$$

where $\boldsymbol{\beta}_\tau$ is the unknown regression parameter. Model (1.1) allows the predictors to have different effects on long-term and short-term risk, and includes the proportional odds model as a special case. Under (1.1), Uno et al. (2007) showed that IPW estimator $\widetilde{\boldsymbol{\beta}}_\tau$, the solution to

$$\widetilde{\mathbf{U}}_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \widehat{w}_i \mathbf{X}_i \left\{ Y_{\tau i} - g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) \right\}, \tag{1.2}$$

is a consistent and asymptotically normal estimator of $\bar{\boldsymbol{\beta}}_\tau$, where $\bar{\boldsymbol{\beta}}_\tau$ is the solution to

$$\mathbf{U}_0(\boldsymbol{\beta}) \equiv E\left[\mathbf{X}\left\{Y_\tau - g(\boldsymbol{\beta}^\intercal \mathbf{X})\right\}\right] = 0,$$

$\widehat{w}_i = \{I(T_i \leq \tau)\delta_i + I(T_i > \tau)\}/\widehat{G}(T_i \wedge \tau)$ and $\widehat{G}(\cdot)$ is the Kaplan-Meier (KM) estimator of $G(\cdot)$. Here, $\bar{\boldsymbol{\beta}}_\tau$ does not depend on the censoring distribution regardless of the adequacy of model 1.1.

To improve the efficiency of $\widetilde{\boldsymbol{\beta}}_\tau$ in the presence of model misspecification, we propose a two-step, imputation-based augmentation procedure that aims to make use of all subjects and incorporate additional information on $\mathbf{S}$. In step I, we fit a flexible

working model, $P(Y_\tau = 1 \mid \mathbf{Z}) = g\{\boldsymbol{\theta}_\tau^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z})\}$, to impute Y_τ as $g\{\widehat{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z})\}$, where $\boldsymbol{\Phi}(\mathbf{Z})$ is a finite set of basis functions of $\mathbf{Z}$ chosen such that $\mathbf{X}$ is in its linear span. Here, $\widehat{\boldsymbol{\theta}}_\tau$ is the minimizer of

$$\widehat{R}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\widehat{w}_i\ell\left\{Y_{\tau i}, g(\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\Phi}_i)\right\} + \lambda_n\mathcal{Q}(|\boldsymbol{\theta}_{[-1]}|), \quad \text{with } 0 \le \lambda_n = o(n^{-\frac{1}{2}})$$

where $\ell(y,x) = y\log\{g(x)\} + (1-y)\log\{1-g(x)\}$, $\boldsymbol{\Phi}_i = \boldsymbol{\Phi}(\mathbf{Z}_i)$, and $\mathcal{Q}(\cdot)$ is a penalty function such as the ridge or LASSO (Friedman et al., 2001). The tuning parameter λ_n controls the amount of regularization and is chosen via cross validation, but restricted to a specified range to ensure the desired rate. In step II, we estimate $\boldsymbol{\beta}_\tau$ as $\widehat{\boldsymbol{\beta}}_\tau$, the solution to

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\left\{g(\widehat{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}_i) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\right\}.$$

In Appendix A.2, we show that $\widehat{\boldsymbol{\beta}}_\tau$ is a consistent estimate of $\bar{\boldsymbol{\beta}}_\tau$, even when $\mathbf{S}$ is measured post-baseline. In addition, we show in Appendix A.3 that $\widetilde{\mathbb{W}}_\tau \equiv n^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\eta}_{\text{IPW},i} + o_p(1)$ and $\widehat{\mathbb{W}}_\tau \equiv n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\eta}_{\text{AUG},i} + o_p(1)$, where

$$\boldsymbol{\eta}_{\text{IPW},i} = \mathbf{F}_{1i} + (w_i - 1)\mathbf{F}_{1i} - \int_0^\tau \psi_i(s)\boldsymbol{\mu}_{F_1}(ds), \quad \mathbf{F}_{1i} = \mathbb{J}^{-1}\mathbf{X}_i\left\{Y_{\tau i} - g(\bar{\boldsymbol{\beta}}_\tau^{\mathsf{T}}\mathbf{X}_i)\right\}$$
$$\boldsymbol{\eta}_{\text{AUG},i} = \mathbf{F}_{1i} + (w_i - 1)\mathbf{F}_{2i} - \int_0^\tau \psi_i(s)\boldsymbol{\mu}_{F_2}(ds), \quad \mathbf{F}_{2i} = \mathbb{J}^{-1}\mathbf{X}_i\left\{Y_{\tau i} - g(\bar{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}_i)\right\},$$

$w_i = \{I(T_i \le \tau)\delta_i + I(T_i > \tau)\}/G(T_i \wedge \tau)$, $\psi_i(s) = \int_0^s \pi(u)^{-1}dM_{ci}(u)$, $\pi(u) = P(T_i > u)$, $M_{ci}(u) = I(T_i \le u, \delta_i = 0) - \int_0^u I(T_i > v)d\Lambda_c(u)$, $\Lambda_c(u) = -\log\{G(u)\}$, $\mathbb{J} = E[\mathbf{X}\mathbf{X}^{\mathsf{T}}\dot{g}(\bar{\boldsymbol{\beta}}_\tau^{\mathsf{T}}\mathbf{X})]$, and $\boldsymbol{\mu}_{F_k}(s) = E\{\mathbf{F}_{ki}I(T_i^\dagger > s)\} = -E\{\mathbf{F}_{ki}I(T_i^\dagger \le s)\}$ for $k = 1, 2$. It follows that $\widehat{\mathbb{W}}_\tau$ and $\widetilde{\mathbb{W}}_\tau$ converge in distribution to zero-mean multivariate normals with covariance matrices $\boldsymbol{\Sigma}_{\text{AUG}} = E(\boldsymbol{\eta}_{\text{AUG},i}^{\otimes 2})$ and $\boldsymbol{\Sigma}_{\text{IPW}} = E(\boldsymbol{\eta}_{\text{IPW},i}^{\otimes 2})$, respectively. It is also shown in **??** that the variance reduction $\Delta\text{var} = \boldsymbol{\Sigma}_{\text{AUG}} - \boldsymbol{\Sigma}_{\text{IPW}}$ takes the form

$$\int_0^\tau \{\text{var}(\mathbf{F}_{1i}|T_i^\dagger > s) - \text{var}(\mathbf{F}_{2i}|T_i^\dagger > s)\}\frac{S(s)^2 d\Lambda_c(s)}{\pi(s)} + \left\{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_2}(\tau)^{\otimes 2}\right\}\int_0^\tau \frac{d\Lambda_c(s)}{\pi(s)}.$$

Although $\widehat{\boldsymbol{\beta}}_\tau$ is often more efficient than $\widetilde{\boldsymbol{\beta}}_\tau$ since the imputation leverages potential non-linear effects and information on $\mathbf{S}$, $\widehat{\boldsymbol{\beta}}_\tau$ is not fully efficient, especially when the imputation model $P(Y_\tau = 1 \mid \mathbf{Z}) = g\{\boldsymbol{\theta}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z})\}$ is misspecified. This motivates us to propose our final estimator as the optimal linear combination of the two candidate estimators, $\widetilde{\boldsymbol{\beta}}_\tau$ and $\widehat{\boldsymbol{\beta}}_\tau$. For simplicity, we only consider component-wise optimal linear combinations and construct $\widehat{\beta}_{\tau j}^{\text{comb}} = \widehat{\mathbf{C}}_j^\intercal(\widetilde{\beta}_j, \widehat{\beta}_j)^\intercal$, where $\widehat{\mathbf{C}}_j = \mathbf{1}^\intercal \widehat{\boldsymbol{\Sigma}}_j^{-1} / \mathbf{1}^\intercal \widehat{\boldsymbol{\Sigma}}_j^{-1} \mathbf{1}$ and $\widehat{\boldsymbol{\Sigma}}_j$ is a consistent estimator for the covariance matrix of $(\widetilde{\mathbb{W}}_{\tau j}, \widehat{\mathbb{W}}_{\tau j})^\intercal$. It is straightforward to see that $\widehat{\beta}_{\tau j}^{\text{comb}}$ is always at least as efficient as $\widetilde{\beta}_{\tau j}$ or $\widehat{\beta}_{\tau j}$. Such a linear combination is desirable because $\widetilde{\beta}_{\tau j} - \widehat{\beta}_{\tau j}$ is not always independent of $\widehat{\beta}_{\tau j}$ and as such, the combined estimator would be more efficient than $\widehat{\beta}_{\tau j}$.

1.2.2 Inference via Resampling

To obtain the combined estimator as well as its corresponding variance, we propose a perturbation resampling procedure similar to those proposed in Uno et al. (2007) for approximating the distribution of $\widetilde{\mathbb{W}}_\tau$. Let $\mathbf{V} = (V_1, ..., V_n)^\intercal$ be an $n \times 1$ vector of independent and identically distributed random variables with mean 1 and variance 1, generated independent of $\mathscr{D}$. Then we obtain a perturbed version of $\widehat{\boldsymbol{\beta}}_\tau$ as $\widehat{\boldsymbol{\beta}}_\tau^*$, the solution to

$$\widehat{\mathbf{U}}_n^*(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} V_i \mathbf{X}_i \{ g(\boldsymbol{\Phi}_i^\intercal \widehat{\boldsymbol{\theta}}^*) - g(\boldsymbol{\beta}^\intercal \mathbf{X}_i) \} = 0$$

where $\widehat{\boldsymbol{\theta}}_\tau^*$ is the minimizer of $\widehat{R}_n^*(\boldsymbol{\theta}) \equiv n^{-1} \sum_{i=1}^n \widehat{w}_i^* \ell\{Y_{\tau i}, g(\boldsymbol{\theta}^\intercal \boldsymbol{\Phi}_i)\} + \lambda_n \mathcal{Q}(|\boldsymbol{\theta}_{[-1]}|)$, $\widehat{w}_i^* = V_i\{I(T_i \le \tau)\delta_i + I(T_i > \tau)\} / \widehat{G}^*(T_i \wedge \tau)$,

$$\widehat{G}^*(t) = \widehat{G}(t) \exp\left\{ -n^{-1} \sum_{i=1}^{n} \int_0^t \frac{(V_i - 1) d\widehat{M}_{ci}(s)}{\widehat{\pi}(s)} \right\},$$

$\widehat{M}_{ci}(t) = I(T_i \le t, \delta_i = 0) + \int_0^t I(T_i > s) d \log\{\widehat{G}(s)\}$ and $\widehat{\pi}(s) = n^{-1} \sum_{i=1}^n I(T_i > s)$. The tuning parameter λ_n can be either reselected for each perturbation or fixed as the initial λ_n selected for the observed data. The latter approach substantially reduces the

computational cost and appears to perform well in our simulation studies. However, when computation is not a concern, we recommend reselecting the optimal tuning parameter for each perturbed sample to ensure the performance of the combined estimator.

In practice, one can generate a large number, say B, of random samples of $\mathbf{V}$ to obtain B realizations of $\widetilde{\boldsymbol{\beta}}_\tau^*$ and $\widehat{\boldsymbol{\beta}}_\tau^*$. Then the unconditional distribution of $(\widetilde{\mathbb{W}}_\tau^\intercal, \widehat{\mathbb{W}}_\tau^\intercal)^\intercal$ can be approximated by the empirical distribution of $\{n^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}_\tau^* - \widetilde{\boldsymbol{\beta}}_\tau), n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_\tau^* - \widehat{\boldsymbol{\beta}}_\tau)\}^\intercal$, conditional on $\mathscr{D}$. This joint distribution can be subsequently used to construct $\widehat{\boldsymbol{\beta}}_\tau^{\text{comb}}$ and the confidence intervals (CIs) for the combined estimator. However, in the realistic situation where sample size is limited, $\widehat{\boldsymbol{\Sigma}}_j^{-1}$ could be unstable due to high correlation between $\widetilde{\boldsymbol{\beta}}_\tau$ and $\widehat{\boldsymbol{\beta}}_\tau$, in which case $\text{var}(\widehat{\boldsymbol{\beta}}_j^{\text{comb}})$ is underestimated. One solution is to use $(\widehat{\boldsymbol{\Sigma}}_j + \epsilon\mathbf{I})^{-1}$ instead of $\widehat{\boldsymbol{\Sigma}}_j^{-1}$, where ϵ is a small number that can be chosen in the order of $O(n^{-\frac{1}{2}})$. For instance, we use $\epsilon = \{\hat{\text{var}}(\tilde{\beta}) + \hat{\text{var}}(\hat{\beta})\}/(2n^{1/2})$ in the example illustrated in Section 4. The standard error (SE) estimates and CIs for various parameters can be constructed empirically based on these resampled realizations. Note that the perturbed scheme is similar to standard bootstrap if $\mathbf{V}$ is generated from a multinomial distribution. However, in the case of rare events, bootstrap is likely to suffer from instability since the subsampling might lead to bootstrap samples with too few events, while the proposed perturbation scheme avoids such a problem by assigning positive weights to all observations.

1.2.3 Improving Estimation of Individualized Treatment Effect (ITE)

When interest lies in identifying individualized treatment rules (ITRs), we assume that data for analysis arise from randomized clinical trials with two treatment groups indexed by $A \in \{0, 1\}$. We use $Y_\tau^{(a)}$ to denote the counterfactual τ-year survival status if a patient is assigned to treatment $A = a$. The optimal binary ITR $\mathscr{I}(\mathbf{X}) : \mathbf{X} \rightarrow$

$\{0,1\}$ maximizing the population average *value function*, $\mathbb{V}(\mathscr{I}) = 1 - E\{\mathscr{I}(\mathbf{X})Y^{(1)} + (1-\mathscr{I}(\mathbf{X}))Y^{(0)}\}$ is the Bayes rule $I\{D(\mathbf{X}) \leq 0\}$, where $D(\mathbf{X}) = E(Y^{(1)}|\mathbf{X}) - E(Y^{(0)}|\mathbf{X})$ (Matsouaka et al., 2014). Under working models $E(Y_\tau^{(a)} \mid \mathbf{X}) = g(\boldsymbol{\beta}_{\tau,a}^\intercal \mathbf{X})$, for any given $\mathbf{x}$ that is a realization of $\mathbf{X}$, we may approximate $D(\mathbf{x})$ by $\bar{D}(\mathbf{x}) = g(\bar{\boldsymbol{\beta}}_{\tau,1}^\intercal \mathbf{x}) - g(\bar{\boldsymbol{\beta}}_{\tau,0}^\intercal \mathbf{x})$, where $\bar{\boldsymbol{\beta}}_{\tau,a}$ is the solution to $E[\mathbf{X}\{Y_\tau^{(a)} - g(\boldsymbol{\beta}^\intercal \mathbf{X})\}] = 0$. Our proposed method could be used to obtain more efficient estimation of $\bar{\boldsymbol{\beta}}_{\tau,a}$, and subsequently a more efficient estimator of $\bar{D}(\mathbf{X})$.

1.2.4 Incorporating Moderate p

When p is not small relative to the number of events, one may obtain a regularized IPW estimator of $\boldsymbol{\beta}_\tau$. To this end, we note that $\widetilde{\boldsymbol{\beta}}_\tau$ and the augmented estimator $\widehat{\boldsymbol{\beta}}_\tau$ are the respective minimizers of $\widetilde{L}_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \widehat{w}_i \ell(Y_{\tau i}, \boldsymbol{\beta}^\intercal \mathbf{X}_i)$ and $\widehat{L}_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \ell\{g(\widehat{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}_i), \boldsymbol{\beta}^\intercal \mathbf{X}_i\}$. Then we estimate $\boldsymbol{\beta}_\tau$ as $\widetilde{\mathscr{B}}_\tau$, the minimizer of the adaptive LASSO (Zhang and Lu, 2007) penalized objective function,

$$\widetilde{L}_n(\boldsymbol{\beta}) + \widetilde{\nu}_n \sum_{j=2}^{p} \left| \beta_j / \widetilde{\beta}_{\tau j} \right|$$

where $\widetilde{\nu}_n \geq 0$ controls the amount of regularization and $\widetilde{\nu}_n \to 0$ as $n \to \infty$. Using similar arguments as given in Zou (2006), one may show that $\widetilde{\mathscr{B}}_\tau \to \bar{\boldsymbol{\beta}}_\tau$ in probability. Similarly, we may obtain a regularized version of $\widehat{\boldsymbol{\beta}}_\tau$, $\widehat{\mathscr{B}}_\tau$, as the minimizer of $\widehat{L}_n(\boldsymbol{\beta}) + \widehat{\nu}_n \sum_{j=2}^p |\beta_j / \widehat{\beta}_{\tau j}|$ for some properly chosen tuning parameter $\widehat{\nu}_n \geq 0$. In Appendix A.4, we show that $n^{\frac{1}{2}}(\widetilde{\mathscr{B}}_\tau - \bar{\boldsymbol{\beta}}_\tau)$ and $n^{\frac{1}{2}}(\widehat{\mathscr{B}}_\tau - \bar{\boldsymbol{\beta}}_\tau)$ converge in distribution to zero-mean multivariate normals. To assess the variability of $\widetilde{\mathscr{B}}_\tau$ and $\widehat{\mathscr{B}}_\tau$, similar perturbation resampling procedures can be employed.

1.3 Simulations

We conducted extensive simulation studies with sample size $n = 500$ to examine the finite sample properties of the proposed estimation procedures and compare them

to existing methods. For the enriched imputation model, we used natural spline bases with 3 knots for each of the covariates across all settings. For each dataset, we used 500 perturbations for variance estimation. Efficiency gain is defined as the relative mean square error minus one in percentage scale. All results are based on 500 simulated datasets for each configuration. We first considered independent censoring in Sections 1.3.1-1.3.3 and generated C from an exponential with mean λ, chosen to achieve designed event rates. In Section 1.3.4, we investigated the case with C dependent on $\mathbf{X}$, and the robustness of various procedures with respect to estimation and prediction. Note that the effective sample size is much smaller than 500 due to the relatively low event rates.

For comparison, we also obtained (i) IPW estimators with censoring weights calculated from KM estimator (IPW_{KM}) as in Uno et al. (2007); (ii) IPW estimators with censoring weights estimated by fitting a Cox model with both $\mathbf{X}$ and S (IPW_{cox}) as in Scharfstein et al. (1999); (iii) an AIPW estimator constructed as in DiRienzo (2009) with censoring weights estimated by KM (AIPW_{KM}); and (iv) an AIPW approach similar to (iii) with the censoring weights calculated using a Cox model (AIPW_{cox}). For our methods, we were also interested in evaluating the added value of S in improving efficiency, and thus constructed the augmented estimators with the imputation step using (I) $\mathbf{X}$ only (AUG_X); and (II) both $\mathbf{X}$ and S ($\text{AUG}_{X,S}$). Finally, we also considered the proposed combined estimator (AUG_{CMB}) that optimally combines IPW_{KM} and $\text{AUG}_{X,S}$.

1.3.1 Low Dimensional Baseline Covariates

We first considered a small $p = 4$ and generated $\mathbf{X}_{-1}$, $T^\dagger$ and S from

$$
(\mathscr{M}_1): \begin{array}{l} \mathbf{X}_{-1} = (X_2, X_3, X_4)^\intercal \sim N(\mathbf{0}, 0.3 + 0.7\mathbb{I}_3), \quad \text{where } \mathbb{I}_d \text{ is the } d \times d \text{ diagonal matrix,} \\ \log(T^\dagger) = 0.5(X_2 + X_3 + X_4) + 0.5X_2^2 + X_3^2 + 0.5X_4^2 - 3 + \text{logit}(U) + \log(\alpha), \\ S = \text{logit}(U) + 0.1X_2 + 0.1X_3 + \sigma_s\varepsilon, \quad \text{with } U \sim \text{Uniform(0,1) and } \varepsilon \sim N(0,1), \end{array}
$$

We let $\sigma_s = 0.5$ or 2 to induce high or moderate correlation between S and $\log(T^\dagger)$, respectively. We generated C from $\mathscr{C}_{\text{KM}} \sim \exp(\lambda)$, an exponential distribution with

Table 1.1: Empirical bias, SE (ESE), average of the estimated SE (ASE) and coverage probabilities (CovP) of the 95% CIs for the low-dimensional setting. Shown also are the percent of efficiency gain (%EffG) relative to the IPW_{KM} estimator.

σ_s		Bias × 100				ESE_{ASE}× 100				CovP				%EffG			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Low Event Rate																	
	IPW_{KM}	0.8	-0.7	-1.1	-1.4	$15.6_{16.2}$	$20.0_{20.0}$	$20.5_{19.6}$	$19.8_{20.1}$	97	96	95	96	-	-	-	-
	AUG_X	1.7	-1.3	-2.0	-0.8	$16.0_{16.0}$	$17.2_{17.2}$	$16.2_{15.6}$	$16.4_{17.3}$	96	96	95	97	-5.9	35.4	57.7	47.0
	IPW_{cox}	-0.9	-1.0	-1.1	-0.9	$14.9_{17.4}$	$18.6_{18.4}$	$18.2_{17.5}$	$18.3_{18.6}$	98	95	96	95	10.3	16.2	26.3	18.3
	AIPW_{KM}	1.9	0.8	-1.0	-0.5	$15.9_{16.9}$	$19.3_{19.5}$	$20.9_{20.3}$	$19.3_{19.7}$	96	95	93	95	-4.0	7.4	-4.1	6.4
0.5	AIPW_{cox}	0.8	0.6	-1.0	-0.5	$15.8_{16.8}$	$18.6_{19.0}$	$19.7_{19.4}$	$18.6_{19.3}$	96	94	94	95	-2.3	15.4	8.1	14.1
	$\text{AUG}_{X,S}$	2.3	-0.7	-2.8	-0.9	$13.5_{13.2}$	$14.2_{13.6}$	$14.5_{13.0}$	$13.9_{13.6}$	95	93	93	95	30.3	98.0	92.0	103.5
	AUG_{CMB}	1.8	-0.2	-2.8	-0.7	$13.5_{13.2}$	$14.0_{13.4}$	$14.4_{12.9}$	$13.9_{13.6}$	95	93	93	94	31.3	103.7	96.7	105.8
	IPW_{cox}	-1.1	-0.9	-1.1	-1.0	$15.2_{17.7}$	$18.7_{18.4}$	$17.9_{17.5}$	$18.4_{18.6}$	98	95	96	95	5.3	13.1	26.7	16.2
	AIPW_{KM}	1.3	0.4	-1.5	-0.9	$15.9_{17.1}$	$19.7_{20.6}$	$20.4_{20.8}$	$19.9_{20.9}$	97	96	95	97	-4.1	2.8	-2.4	-1.1
2	AIPW_{cox}	-0.3	-0.0	-1.5	-0.7	$15.8_{16.8}$	$18.8_{19.8}$	$18.8_{19.4}$	$19.0_{20.1}$	96	96	96	96	-2.3	12.1	14.0	9.2
	$\text{AUG}_{X,S}$	2.4	-0.7	-2.7	-0.6	$15.5_{15.3}$	$16.2_{16.0}$	$15.6_{14.8}$	$15.9_{16.1}$	94	95	93	96	-0.3	50.5	63.2	54.8
	AUG_{CMB}	1.4	-0.3	-2.8	-0.2	$15.2_{15.0}$	$15.9_{15.8}$	$15.4_{14.7}$	$15.7_{16.0}$	94	95	94	96	4.7	56.4	67.8	60.2
Moderate Event rate																	
	IPW_{KM}	0.2	0.0	-0.4	-1.0	$11.6_{11.7}$	$14.1_{14.0}$	$13.7_{13.9}$	$14.1_{14.1}$	95	96	96	94	-	-	-	-
	AUG_X	0.4	-0.6	-0.8	-0.1	$11.3_{11.5}$	$12.5_{12.7}$	$11.6_{12.0}$	$12.6_{12.7}$	95	96	96	95	4.3	25.6	39.8	26.6
	IPW_{cox}	-0.4	0.0	-0.2	-0.6	$11.3_{12.2}$	$13.6_{13.2}$	$12.4_{12.9}$	$13.3_{13.3}$	97	94	96	94	4.0	7.0	21.9	13.0
	AIPW_{KM}	0.5	-0.0	-0.7	-0.6	$11.2_{11.7}$	$13.3_{13.6}$	$14.0_{14.1}$	$13.2_{13.7}$	95	95	94	96	6.2	11.7	-3.5	14.5
0.5	AIPW_{cox}	-0.1	0.0	-0.5	-0.4	$11.1_{11.6}$	$13.2_{13.3}$	$13.0_{13.5}$	$12.8_{13.4}$	95	96	96	96	8.9	13.9	11.3	21.9
	$\text{AUG}_{X,S}$	0.5	-0.9	-0.7	0.3	$10.1_{10.3}$	$11.5_{11.1}$	$10.8_{10.7}$	$11.4_{11.1}$	96	94	95	94	30.7	49.4	59.9	54.0
	AUG_{CMB}	0.3	-0.8	-0.7	0.3	$10.1_{10.3}$	$11.4_{11.0}$	$10.8_{10.6}$	$11.4_{11.0}$	96	94	94	94	31.0	50.7	61.2	54.4
	IPW_{cox}	-0.3	0.0	-0.2	-0.5	$11.4_{12.3}$	$13.6_{13.2}$	$12.4_{12.9}$	$13.3_{13.3}$	97	94	96	94	1.9	6.9	22.1	13.3
	AIPW_{KM}	0.5	-0.1	-0.7	-0.7	$11.3_{11.9}$	$13.7_{14.1}$	$13.9_{14.2}$	$13.7_{14.2}$	95	96	95	95	3.4	4.5	-3.6	6.4
2	AIPW_{cox}	-0.2	-0.1	-0.5	-0.4	$11.2_{11.7}$	$13.5_{13.6}$	$12.8_{13.4}$	$13.1_{13.7}$	96	95	97	96	5.7	9.0	14.1	17.1
	$\text{AUG}_{X,S}$	0.7	-0.7	-1.3	0.0	$10.9_{11.2}$	$12.1_{12.2}$	$11.8_{11.6}$	$12.1_{12.2}$	96	95	95	94	10.7	33.6	33.7	36.0
	AUG_{CMB}	0.4	-0.6	-1.3	0.1	$10.9_{11.1}$	$12.2_{12.1}$	$11.8_{11.6}$	$12.2_{12.2}$	95	95	95	95	11.3	33.3	33.7	35.4

mean λ. We considered (i) a low event rate (12−18% by τ) and heavy censoring setting (65−74% before τ) with $\{\alpha = 12, \lambda = 0.5\}$; and (ii) a moderate event rate (25−34% by τ) and moderate censoring setting (37−50% before τ) with $\{\alpha = 6, \lambda = 1\}$, where $\tau = 0.8$.

As shown in Table 1.1, estimators of $\boldsymbol{\beta}_\tau$ have negligible biases across all settings. Our proposed estimators AUG_X, $\text{AUG}_{X,S}$ and AUG_{CMB} perform substantially better than all existing estimators with efficiency gains relative to IPW_{KM} as high as 106%, while existing augmented estimators such as AIPW_{cox} only have very modest gains. In general, $\text{AUG}_{X,S}$ performs better than AUG_X, which indicates that incorporating S can lead to additional efficiency gains. The resampling-based inference procedures also perform well with the SE estimates close to the empirical SE and the empirical coverage probability (CovP) of the 95% CIs close to their nominal level.

1.3.2 ITE Estimation

We also studied the performance of our proposed estimators for ITE under a randomized clinical trial setting with treatment group $A \in \{0,1\}$ and $n_a = 500$ for $a = 0, 1$. We generated $\mathbf{X}_{-1} = (X_2, X_3, X_4)^\intercal$ from $N(\mathbf{0}, 0.3 + 0.7\mathbb{I}_3)$ for both groups, and $T^\dagger$ from $\log(T^\dagger) = (0.1 + 0.1A)X_2 + (0.4 - 0.3A)X_3 + (0.1 + 0.2A)X_4 + 0.5X_2^2 + X_3^2 + 0.5X_4^2 - 3 + \text{logit}(U) + \log(12 - 4A)$ with $U \sim$ Uniform(0,1). The surrogate variable $S = \text{logit}(U) + 0.5A + 0.1X_2 + 0.1X_3 + 0.3\varepsilon$, where $\varepsilon \sim N(0,1)$ and the correlation between S and $\log(T^\dagger)$ was around 60% within group. We generated $C \mid A$ from an exponential distribution with mean 0.6 leading to event rates by τ of about 16-27% for $A = 1$ and 12-21% for $A = 0$. We obtained estimates of $\bar{\boldsymbol{\beta}}_{\tau a}$ for $\tau = 0.9$ using various methods and then estimated the ITE score $\bar{D}(\mathbf{x}) = g(\bar{\boldsymbol{\beta}}_{\tau,1}^\intercal \mathbf{x}) - g(\bar{\boldsymbol{\beta}}_{\tau,0}^\intercal \mathbf{x})$ for two specific covariate levels: (i) $\mathbf{x} = (1,1,1,1)^\intercal$; and (ii) $\mathbf{x} = (1,-0.5,-0.5,-0.5)^\intercal$.

Table 1.2: Empirical bias, standard errors (ESE), average of the estimated SE (ASE) and coverage probabilities (CovP) of the 95% confidence intervals for covariate-specific risk, $\pi_a(\mathbf{x}) = g(\boldsymbol{\beta}_{\tau,a}^\intercal \mathbf{x})$ within each treatment group $A = a$ and the ITE $D(\mathbf{x})$. Shown also are the percent of efficiency gain (%EffG) relative to the IPW_{KM} estimator. All numbers have been multiplied by 100.

	$\pi_0(\mathbf{x})$				$\pi_1(\mathbf{x})$				$D(\mathbf{x})$			
	Bias	ESE_{ASE}	CovP(%)	%EffG	Bias	ESE_{ASE}	CovP(%)	%EffG	Bias	ESE_{ASE}	CovP(%)	%EffG
Example 1: $\mathbf{x} = (1,1,1,1)^\intercal$												
IPW_{KM}	0.35	$5.37_{5.36}$	94	-	0.31	$4.61_{4.68}$	96	-	0.04	$6.96_{7.17}$	96	-
IPW_{cox}	-0.02	$4.72_{4.99}$	96	30.16	0.02	$4.13_{4.38}$	95	25.29	-0.04	$6.15_{6.68}$	97	28.36
AIPW_{KM}	0.81	$5.90_{5.68}$	94	-18.42	0.51	$4.86_{5.04}$	95	-10.29	0.30	$7.35_{7.63}$	96	-10.38
AIPW_{cox}	0.49	$5.40_{5.27}$	94	-1.65	0.25	$4.53_{4.72}$	95	3.72	0.24	$6.78_{7.12}$	96	5.24
AUG_X	0.05	$4.37_{4.44}$	96	51.30	0.05	$4.02_{4.00}$	93	32.12	-0.00	$5.76_{6.01}$	96	46.44
$\text{AUG}_{X,S}$	0.19	$4.20_{3.98}$	93	64.19	0.02	$3.81_{3.60}$	93	47.19	0.17	$5.51_{5.38}$	95	59.41
AUG_{CMB}	0.15	$4.17_{3.93}$	93	66.27	-0.01	$3.75_{3.54}$	93	52.17	0.16	$5.46_{5.34}$	95	62.86
Example 2: $\mathbf{x} = (1,-0.5,-0.5,-0.5)^\intercal$												
IPW_{KM}	0.52	$4.57_{4.40}$	94	-	0.45	$3.96_{3.98}$	95	0.00	0.07	$6.00_{5.95}$	95	0.00
IPW_{cox}	0.15	$4.18_{4.46}$	95	20.83	0.23	$3.82_{4.02}$	96	8.30	-0.08	$5.68_{6.02}$	97	11.60
AIPW_{KM}	0.46	$4.56_{4.53}$	95	0.78	0.56	$4.04_{4.14}$	94	-4.58	-0.10	$6.16_{6.15}$	95	-5.07
AIPW_{cox}	0.19	$4.29_{4.32}$	95	14.57	0.34	$4.04_{4.00}$	93	-3.55	-0.16	$5.98_{5.90}$	95	0.62
AUG_X	0.76	$4.06_{3.96}$	94	23.96	0.71	$3.78_{3.67}$	93	7.32	0.06	$5.64_{5.41}$	94	13.41
$\text{AUG}_{X,S}$	0.81	$3.77_{3.50}$	93	42.09	0.90	$3.56_{3.29}$	92	17.71	-0.09	$5.32_{4.82}$	92	27.45
AUG_{CMB}	0.73	$3.72_{3.47}$	93	46.78	0.83	$3.53_{3.25}$	92	20.48	-0.10	$5.26_{4.78}$	93	30.04

In Table 1.2, we compared the model-based estimate of conditional τ-year risk for patients treated with either $A = 1$ or $A = 0$ and the ITE score $D(\mathbf{x})$ using regression coefficients obtained from various methods. Again, while all estimators have negligible bias, the proposed methods are substantially more efficient than existing methods.

The proposed resampling procedures also work well for making inference with proper SE estimation and empirical coverage for the CIs.

1.3.3 Baseline Model Regularization

We also conducted simulation studies to assess the finite sample performance of the proposed methods that incorporate regularization with $p = 11$. We generated $\mathbf{X}_{-1} \sim N(\mathbf{0}, \mathbb{I}_{10})$ and

$$\log(T^\dagger) = X_2 + X_3 + X_4 + 0.5X_2^2 + X_3^2 + 0.5X_4^2 - 3 + \text{logit}(U) + \log(6).$$

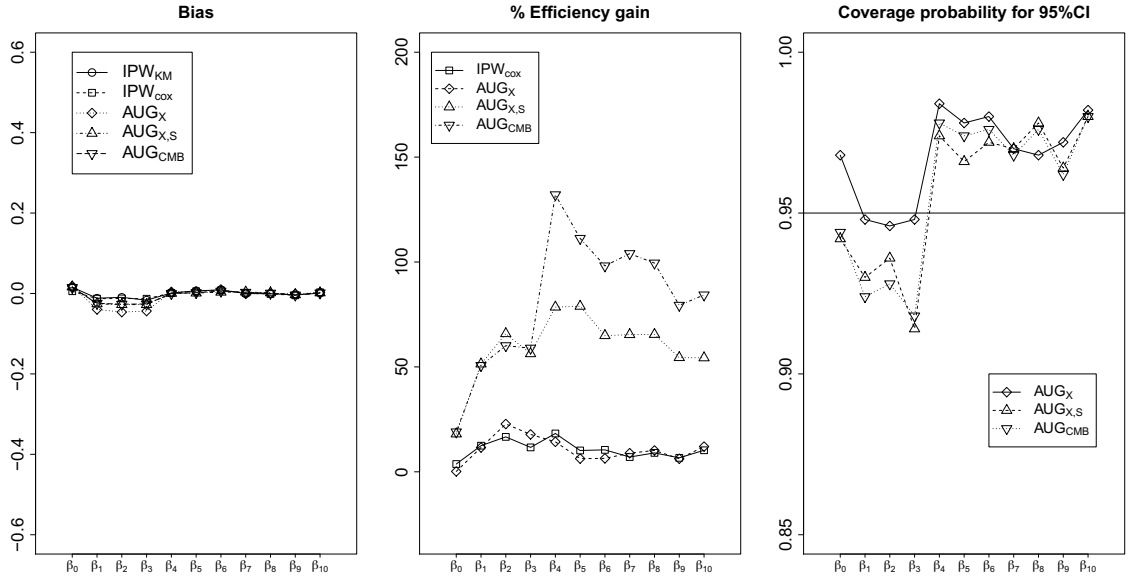
We generated C and S as in Section 1.3.1 and let $\lambda = 1$, $\sigma_s = 0.5$ and $\tau = 1$, leading to an observed event rate by τ of about $26 - 38\%$, and a correlation between S and $\log(T^\dagger)$ of about $0.55 - 0.7$. The effective sample size is around 100-200, not large relative to $p = 11$. To control overfitting in the imputation model, L_1 penalized logistic regression (Friedman et al., 2010) with IPW was employed when estimating the imputation model.

Figure 1.1 summarizes the results for our estimators, as well as IPW_{KM} and IPW_{cox} as benchmarks for efficiency assessment. The AIPW methods were not included as no associated regularization procedures were available. The proposed methods that incorporate both $\mathbf{X}$ and S are generally more efficient than IPW_{cox} and IPW_{KM}, as in the previous settings. The resampling procedures also perform well with empirical CovP ranging from 92-95% for informative signals. The CovPs for the zero signals range from 96%-98%, which is expected owing to the oracle properties.

1.3.4 Dependent censoring

We next investigated the robustness of various estimators with respect to the independent censoring assumption. We focused on $\widehat{\boldsymbol{\beta}}_\tau^{\text{comb}}$ and a variation of our estimator with IPW weights estimated from a Cox model, denoted by $\text{AUG}_{\text{CMB}}^{\text{cox}}$. We let $p = 4$, generate $\mathbf{X}_{-1}$, S, $T^\dagger$ from $(\mathscr{M}_1)$ with $\alpha = 12$ and $\mathscr{C}_{\text{KM}} \sim exp(\lambda)$ with $\lambda = 0.5$, and

Figure 1.1: Simulation results for regularized baseline models



then generated C from either

($\mathscr{C}_{\text{cox}}$): a Cox model with $C = \mathscr{C}_{\text{KM}} \exp(\boldsymbol{\alpha}_1^{\intercal}\mathbf{X}_{-1})$; or

($\mathscr{C}_{\text{ncox}}$): a non-Cox model with $C = \mathscr{C}_{\text{KM}} + \boldsymbol{\alpha}_2^{\intercal} \exp(-\mathbf{X}_{-1})$

where $\boldsymbol{\alpha}_1 = (-0.2, 0, 0)^{\intercal}$ and $\boldsymbol{\alpha}_2 = (0.2, 0, 0)^{\intercal}$. Under such parameterization, censoring is associated with X_2 with a hazard ratio between 1.1 and 1.9. Results are summarized in Web Table 1 for $\tau = 0.8$. When C is from ($\mathscr{C}_{\text{cox}}$), IPW_{cox}, AIPW_{cox}, and $\text{AUG}_{\text{CMB}}^{\text{cox}}$ all have negligible bias with proper coverage levels for the CIs. This is as expected, since these methods calculate IPW weights from fitting a Cox model for C. When C is from ($\mathscr{C}_{\text{ncox}}$), none of these methods provide consistent estimators, but AUG_{CMB} and $\text{AUG}_{\text{CMB}}^{\text{cox}}$ generally have much smaller bias than existing estimators. The estimator AUG_{CMB} is also not very sensitive to the departure of the independent censoring assumption. Compared to the benchmark estimator of IPW with KM weights, our estimators are substantially more efficient with respect to mean square error.

Although τ-GLM is a more flexible model, its associated inference procedures require stronger assumptions regarding the censoring distribution. The traditional Cox model makes stronger model assumptions but requires weaker censoring assumptions. To compare different procedures including the Cox model under model misspecification, we compared the performance of various procedures with respect to the accuracy in predicting τ-year survival based on: (i) the area under the Receiver Operating Characteristic curve (AUC); and (ii) incremental proportion of explained variance (IPEV), defined as $1 - E[\{I(T^\dagger \leq \tau) - \widehat{\mathcal{P}}_\tau(\mathbf{X})\}^2]/E[\{I(T^\dagger \leq \tau) - \widehat{P}_\tau\}^2]$ (Gerds et al., 2008), where $\widehat{\mathcal{P}}_\tau(\cdot)$ is the estimated conditional risk function, and $\widehat{P}_\tau$ is the KM estimator of $P(T^\dagger \leq \tau)$ using training data. For each trained model, the corresponding accuracy parameters are estimated via Monte Carlo using a fully observed independent validation dataset with a sample size of $100,000$. The true underlying data are generated from two misspecified models: $(\mathscr{M}_1)$ as defined previously in this section; and $(\mathscr{M}_2)$ $\mathbf{X}_{-1} = (X_2, X_3)^\intercal$ generated from multivariate log-normal with mean zero, covariance (0.1, 1.2) and correlation -0.1, $\log(T^\dagger) = \log(X_2^2+3X_3^2)+0.1(X_2^2+2X_3^2)\xi+0.5$ and $S = 0.4\log(X_2^2 + X_3^2)\xi + \epsilon$ with $\xi \sim \text{logistic}(0, 1)$, $\epsilon \sim N(0, 1)$. The τ-GLM misspecifies the covariate effects for Model $(\mathscr{M}_1)$, and does not capture the covariate effects or heteroskedasticity correctly for $(\mathscr{M}_2)$. For both $(\mathscr{M}_1)$ and $(\mathscr{M}_2)$, we considered three different censoring distributions $(\mathscr{C}_{\text{KM}})$, $(\mathscr{C}_{\text{cox}})$ and $(\mathscr{C}_{\text{ncox}})$. For Model $(\mathscr{M}_2)$, we let $\lambda = 1$, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = (0.2, 0)^\intercal$, and $\tau = 0.5$, which is about the 70th percentile of the observed follow-up time. For the imputation step of our methods, we used 5-knot natural spline bases. The results are shown in Table A.2. They suggest that our proposed procedures are robust in model misspecification and covariate dependent censoring with respect to prediction performance. Compared to the Cox model, the proposed method achieves similar prediction accuracy under $(\mathscr{M}_1)$, and substantially higher accuracy under $(\mathscr{M}_2)$, across different censoring patterns.

1.4 Example

We illustrate the proposed procedures using a dataset from the AIDS Clinical Trial Group (ACTG) Protocol 175 (Hammer et al., 1996). This study consists of 2467 patients randomized to 4 different treatments; zidovudine (ZDV) only, ZDV+didanosine (ZDV+DDI), ZDV+zalcitabine (ZDV+ZAL), and didanosine (DDI) only. For illustration, we focus on patients treated with ZDV only (mono group, n=619) and ZDV+DDI (combo group, n=612) and use the study-defined secondary endpoint - 'time to progression' defined as occurrence of death or AIDS as defining events. Our goal is to develop (I) risk prediction models for progression by week 144 within each treatment group; and also (II) an ITE prediction model. Baseline covariates, $\mathbf{X}$, include baseline CD4 counts (CD4), age, intravenous drug use (IV), use of ZDV within 30 days prior to randomization (pZDV), and symptomatic status (SS). We include three intermediate outcomes for $\mathbf{S}$: (i) short-term immune response defined as CD4 count at week 24; and (ii) short-term toxicity indicating whether the patient has experienced grade 3 or higher abnormal lab results by week 24; and (iii) short-term tolerability indicating whether the patient has experienced grade 3 or higher signs and symptoms by week 24. Missing values were handled by carrying forward the last observation. Note that less than 0.5% subjects experienced an event of interest and another 3.5% subjects were censored before week 24. These subjects were excluded from the analyses.

We first consider risk prediction within each treatment group based on $\mathbf{X}$ using various methods. For the enriched imputation modeling required by our approach, we use spline bases with 3 knots for all continuous variables. Resampling with 500 replications is used to generate the variance of the IPW_{KM} method and our proposed methods, and bootstrap was used for other methods. As shown in Table 1.3, the point estimates from different methods are comparable to each other. CD4 is a significant risk predictor for both treatment groups, while SS and pZDV are only significant

Table 1.3: Estimated intercept (Int) and covariate effects for the week 144 progression risk prediction models among those treated with ZVD and with ZVD+DDI arms from ACTG175.

	Coefficient$_{SE}$						% EffG					
	Int	IV	Age	pZDV	SS	CD4	Int	IV	Age	pZDV	SS	CD4
ZDV arm: Event rate=15%												
IPW_{KM}	$-0.31_{0.79}$	$0.06_{0.37}$	$0.33_{0.15}$	$0.14_{0.27}$	$-0.23_{0.34}$	$-0.75_{0.16}$						
IPW_{cox}	$-0.59_{0.77}$	$-0.12_{0.37}$	$0.38_{0.15}$	$0.32_{0.26}$	$-0.30_{0.34}$	$-0.75_{0.16}$	3.1	0.2	2.9	3.5	1.0	3.5
AIPW_{KM}	$-0.28_{0.81}$	$0.07_{0.41}$	$0.29_{0.15}$	$0.22_{0.28}$	$-0.20_{0.36}$	$-0.72_{0.16}$	-6.4	-16.8	-4.6	-10.5	-10.7	9.1
AIPW_{cox}	$-0.63_{0.81}$	$-0.11_{0.42}$	$0.35_{0.15}$	$0.40_{0.28}$	$-0.27_{0.36}$	$-0.71_{0.16}$	-6.2	-19.9	-0.3	-8.7	-11.3	6.5
AUG_{X}	$-0.78_{0.63}$	$0.06_{0.27}$	$0.28_{0.12}$	$0.10_{0.19}$	$-0.11_{0.24}$	$-0.55_{0.13}$	54.9	84.7	46.4	103.1	94.4	48.0
$\text{AUG}_{X,S}$	$-0.28_{0.66}$	$0.06_{0.28}$	$0.27_{0.13}$	$0.18_{0.20}$	$-0.15_{0.25}$	$-0.69_{0.14}$	43.2	74.9	40.7	86.9	86.9	39.3
AUG_{CMB}	$-0.27_{0.67}$	$0.06_{0.27}$	$0.23_{0.13}$	$0.20_{0.19}$	$-0.11_{0.24}$	$-0.69_{0.14}$	39.3	84.5	40.8	101.1	104.6	32.0
ZDV+DDI arm: Event rate=12%												
IPW_{KM}	$-0.69_{0.89}$	$0.31_{0.42}$	$0.21_{0.19}$	$0.49_{0.33}$	$1.22_{0.34}$	$-0.85_{0.20}$						
IPW_{cox}	$-0.96_{0.90}$	$0.16_{0.42}$	$0.22_{0.19}$	$0.71_{0.33}$	$1.27_{0.33}$	$-0.81_{0.20}$	-0.6	-0.3	3.5	1.6	2.8	-1.8
AIPW_{KM}	$-0.56_{0.96}$	$0.33_{0.48}$	$0.20_{0.19}$	$0.50_{0.37}$	$1.19_{0.34}$	$-0.86_{0.21}$	-12.3	-20.7	0.7	-16.5	-1.1	-15.1
AIPW_{cox}	$-0.87_{0.96}$	$0.19_{0.48}$	$0.21_{0.19}$	$0.72_{0.37}$	$1.27_{0.34}$	$-0.82_{0.21}$	-13.5	-20.9	1.5	-19.3	-0.1	-16.9
AUG_{X}	$-0.62_{0.78}$	$0.33_{0.35}$	$0.18_{0.16}$	$0.50_{0.29}$	$1.19_{0.33}$	$-0.83_{0.21}$	33.0	45.5	43.1	33.3	3.0	-14.8
$\text{AUG}_{X,S}$	$-0.94_{0.70}$	$0.21_{0.31}$	$0.19_{0.15}$	$0.34_{0.24}$	$0.88_{0.32}$	$-0.64_{0.18}$	64.9	85.9	70.3	91.1	13.6	24.3
AUG_{CMB}	$-1.04_{0.70}$	$0.15_{0.30}$	$0.17_{0.14}$	$0.30_{0.24}$	$0.97_{0.32}$	$-0.69_{0.18}$	61.9	94.8	79.9	88.0	11.2	20.1

for those treated by ZDV+DDI. In general, our estimators are substantially more efficient than alternative methods. For the ZDV group, the efficiency gain appears to come primarily from augmentation from $\mathbf{X}$ alone. The intermediate outcomes do not appear to be highly predictive of the progression, and consequently, incorporating $\mathbf{S}$ in the augmentation resulted in moderate efficiency loss due to overfitting. On the other hand, for the ZDV+DDI group, incorporating $\mathbf{S}$ substantially improves the efficiency of the estimators. The combined estimator AUG_{CMB} generally performs well.

Based on these risk models, we also derived models for predicting the individualized treatment benefit of ZDV+DDI versus ZDV alone on the progression risk at week 144. In addition to IPW methods, we also added the prediction results from the Cox model. The predicted risks within each treatment group for two examples of covariates, along with their specific treatment differences, are presented in Table 1.4. Example 1 is a 22-year old non-IV drug user who had a baseline CD4 count 100 cells/mm^3 without symptoms at baseline, and did not take ZDV 30 days prior to randomization. The treatment difference for this patient was estimated around -0.18 with SE around 0.12 based on our methods, suggesting a moderate treatment

Table 1.4: Predicted progression risk at week 144 among those treated with ZDV/DDI ($A = 1$) and ZVD ($A = 0$), denoted by $\{\pi_a(\mathbf{x}), a = 0, 1\}$, for two examples of covariates and along with their treatment effects using ACTG 175 study.

	ZDV+DDI			ZDV			Treatment Difference		
	$\pi_1(\mathbf{x})$	SE	%EffG	$\pi_0(\mathbf{x})$	SE	%EffG	$D(\mathbf{x})$	SE	%EffG
Example 1: IV=No, Age=22, pZDV=No, SS=No, CD4=100									
IPW_{KM}	0.26	0.0953	-	0.42	0.1113	-	-0.16	0.15	-
IPW_{cox}	0.22	0.0870	20.00	0.38	0.1071	7.89	-0.16	0.1383	12.95
AIPW_{KM}	0.27	0.1058	-18.91	0.41	0.1112	0.15	-0.14	0.1520	-6.58
AIPW_{cox}	0.23	0.0955	-0.44	0.36	0.1072	7.66	-0.14	0.1425	6.33
Cox	0.14	0.0617	138.38	0.39	0.1094	3.43	-0.24	0.1247	38.79
AUG_X	0.26	0.0789	45.79	0.33	0.0858	68.07	-0.07	0.1168	58.13
$\text{AUG}_{X,S}$	0.24	0.0727	71.69	0.40	0.0929	43.57	-0.17	0.1179	55.33
AUG_{CMB}	0.21	0.0715	77.53	0.39	0.0924	45.01	-0.18	0.1165	59.19
Example 2: IV=Yes, Age=40, pZDV=Yes, SS=No, CD4=350									
IPW_{KM}	0.11	0.0449	-	0.19	0.0531	-	-0.07	0.0692	-
IPW_{cox}	0.11	0.0449	-0.11	0.19	0.0515	6.42	-0.07	0.0683	2.67
AIPW_{KM}	0.12	0.0542	-31.26	0.20	0.0622	-27.02	-0.08	0.0805	-26.19
AIPW_{cox}	0.12	0.0514	-23.58	0.19	0.0612	-24.78	-0.07	0.0779	-21.10
Cox	0.12	0.0465	-6.57	0.19	0.0581	-16.39	-0.06	0.0752	-15.39
AUG_X	0.12	0.0399	26.82	0.20	0.0417	62.46	-0.07	0.0569	47.68
$\text{AUG}_{X,S}$	0.13	0.0375	43.12	0.20	0.0431	51.95	-0.07	0.0561	51.87
AUG_{CMB}	0.09	0.0368	48.60	0.18	0.0412	65.83	-0.09	0.0540	64.27

difference. Note that the proposed estimators provide more efficient estimation in treatment difference than all other methods including the Cox model. The Cox model has a smaller SE in risk probability estimation for the ZDV+DDI arm, possibly due to smaller estimated probability. Example 2 is a 40-year old IV drug user who had a baseline CD4 count 350 cells/mm^3 without symptoms at baseline, and took ZDV 30 days prior to randomization. For this patient, the treatment difference was estimated as -0.09 with SE around 0.05, also suggesting a modest treatment difference. For this patient, all the methods provide similar point estimates and the proposed estimators are substantially more efficient than other methods, including the Cox model.

1.5 Remarks

Leveraging information on post-baseline intermediate variables and allowing for possible model misspecification, our proposed robust IPW estimating procedure improves the efficiency of covariate-specific t-year survival prediction by making use of information from censored subjects. Compared with other methods, our approach has the

following advantages: (i) the consistency of the estimator and the efficiency gain do not rely on the correctly-specified working model, and the degree of efficiency gain is actually even larger under the incorrect specification of the risk model; (ii) by allowing a flexible enriched imputation model that incorporates non-linear effects, the proposed method can achieve higher efficiency gains; and (iii) the incorporation of regularization for settings where p is not small relative to the effective sample size.

The proposed estimator $\widehat{\boldsymbol{\beta}}$ is related to the DiRienzo (2009) AIPW estimator, which is the solution to $n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\{\widehat{Y}_i + \widehat{w}_i(Y_{\tau i} - \widehat{Y}_i) - g(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i)\} = 0$, where $\widehat{Y}_i = g(\widehat{\boldsymbol{\gamma}}^{\intercal}\mathbf{X}_i)$ and $\widehat{\boldsymbol{\gamma}}$ is the solution to $0 = n^{-1}\sum_{i=1}^{n}\{I(T_i \leq \tau)\delta_i + I(T_i > \tau))\mathbf{X}_i(y - g(\boldsymbol{\gamma}_{\tau}^{\intercal}\mathbf{X}_i)\}$. The main differences between $\widehat{\boldsymbol{\beta}}$ and this AIPW estimator is that we impute $Y_{\tau i}$ by a more flexible model instead of $\widehat{Y}_i$, and we also exclude the calibration term $\widehat{w}_i(Y_{\tau i} - \widehat{Y}_i)$. These differences enable us to further improve efficiency and incorporate regularization easily.

Similar to other IPW methods, the consistency of our proposed τ-GLM estimators relies on the assumption that either C is independent of covariates or $C \mid \mathbf{X}$ can be modeled correctly via semi-parametric approaches. When both assumptions fail, our proposed estimators are less sensitive to the violation of such assumptions than the simple IPW estimator, as suggested by our numerical results. It is also important to note that additional distributional assumptions about C are required even if model (1.1) holds for a given τ, due to the curse of dimensionality (Robins and Ritov, 1997). This differs from the standard Cox model, which requires much stronger model assumptions. When both the Cox model and the τ-GLM are misspecified, the proposed estimators have the advantage of being free of censoring, potentially leading to risk models with higher prediction performance.

In our numerical studies, we include spline functions of individual variables for an enriched imputation model. However, other basis functions can be considered provided that X_j is a linear combination of $\boldsymbol{\Phi}(\mathbf{Z})$. In most of our simulations and examples, we only used 3 knots for each variable so that stable estimates could be

obtained for the enriched model. A larger number of basis functions can be used for settings where a larger effective sample size is available, although an overly complex model may lead to substantial overfitting and consequently compromise the efficiency gain of the final augmented estimators in the finite sample. The efficiency gain from incorporating auxiliary covariates would depend on how much additional information they have on the outcome, given targeted baseline covariates. For example, if an intermediate variable is highly correlated with some of the baseline covariates, incorporating it using our method would not lead to substantial efficiency gains. It is thus important to select proper surrogate variables in the rare event clinical trial setting.

Post-baseline covariates cannot be used directly in the working baseline prediction model, but we incorporate them in the estimating process to improve the efficiency. However, incorporating post-baseline covariates needs to be handled with caution. In our example using ACTG 175 data, we incorporated covariates collected at week 24 as intermediate variables when less than 0.5% subjects had events and a negligible number of subjects were censored. In practice, this method could be useful in a setting where the post-baseline auxiliary variable is collected at the very early stage of a study, prior to occurrence of events and censoring. One example could be the collection of short-term treatment response markers. However, if collected at a later stage of a study, an intermediate variable would be missing for a non-negligible portion of subjects who already experienced events or dropped out of the study prior to the measurement time. The proposed methods are not directly applicable to such settings and further research is warranted to incorporate short-term outcomes that are subject to censoring.

Chapter 2

Augmented T-year Survival Regression with Time Dependent Auxiliary Covariates

Abstract

Risk prediction plays an important role in precision medicine. Individualized disease prevention and treatment strategies can be formed optimally according to the predicted risks. In many clinical settings, it is of great interest to develop models for predicting the τ-year risk of developing a clinical event using baseline covariates. Such τ-year risk models can be estimated by fitting a flexible time specific generalized linear model (GLM). However, efficient and robust estimation of the risk model is challenging under heavy censoring and potential model mis-specification. Incorporating intermediate outcome information could potentially improve the efficiency of the prediction model. However, existing augmentation methods largely do not allow intermediate outcomes to be subject to censoring and may yield invalid results under model mis-specification. In this paper, we propose a two-step augmentation method to improve the estimation of τ-year risk model by leveraging longitudinally collected intermediate outcome information that is subject to censoring. Our method allows for easy incorporation of regularization to accommodate moderate covariate size and rare events. We also propose resampling methods to assess the variability of our proposed estimators. Numerical studies show that the proposed point and interval estimation procedures perform well in finite sample. We also demonstrate that our proposed estimators are substantially more efficient compared to existing methods. We also illustrate the proposed methods using data from ACTG175, a randomized clinical trial on HIV-infected subjects.

2.1 Introduction

In clinical practice, it is often of interest to predict patients' time specific survival probabilities from baseline for optimal treatment and monitoring plan. Data from existing randomized clinical trials or observational studies can be used to construct such prediction models. One can pursue the flexible τ-year time specific generalized linear model (τ-GLM) since it allows different factors for short-term and long-term risk estimation (Uno et al., 2007). Compared to traditionally used Cox proportional hazard model , τ-GLM is also more robust to the model mis-specification (Zheng and Cai, 2017), and results in an estimator which does not depend on censoring. However, as an inverse probability weighted (IPW) binary regression model where the weights are zero for subjects who are censored before τ, the regular τ-GLM could suffer inefficiency because the information from these subjects does not contribute to the model estimation. To improve the efficiency of the $\tau-$GLM, one could potentially incorporate the baseline information from those subjects censored before τ and intermediate covariates collected post baseline but before τ into the estimation procedure.

In general, for IPW estimation, constructing the weights as a function of baseline covariates could potentially improve the efficiency of the estimation, even when the censoring is independent of these covariates (Robins et al., 1994). The doubly robust augmented IPW (AIPW) method was proposed to augment the estimating function with the outcome imputation so that the estimation is consistent when either the outcome model or weight model is correctly specified (Scharfstein et al., 1999; Tsiatis, 2006), and it was also implemented by DiRienzo (2009) to τ-GLM. The efficiency of the AIPW estimator is higher than IPW estimator when the outcome model is correctly specified, but it is often difficult to achieve due to the mis-specified model. Zheng and Cai (2017) proposed a two-step imputation based procedure that incorporates auxiliary information including post-baseline non-censored covariates to

improve the efficiency. However, none of above methods is suitable for incorporating those post-baseline covariates that are also impacted by the censoring.

Other methods such as the landmark approach (Gray, 1994; van Houwelingen and Putter, 2011; Parast et al., 2014) could be used to incorporate post-baseline covariates to improve the survival probability prediction, but mostly focus on improving survival probability prediction at a post-baseline time point rather than developing a baseline prediction regression model. Lu and Tsiatis (2008) used an augmentation procedure such that the baseline and post-baseline covariates can be used to improve the efficiency of estimating the log likelihood from a Cox model, however, such an estimator depends on censoring distribution and if the Cox model does not hold, it converges to a parameter that might be difficult to interpret.

To be able to incorporate post-baseline covariates with non-negligible censored values to improve $\tau-$GLM, we propose an imputation method which allows us to incorporate those post-baseline covariates at a pre-specified time point, and with multiple time points, we further develop a systematic approach to construct a linear combination of all estimators to ensure optimal efficiency. The rest of the manuscript is organized as follows. Section 2 introduces the estimating procedure and a resampling method to obtain the inference. Section 3 presents the simulation results showing the consistency and the efficiency gain of the proposed estimator. Section 4 illustrates the proposed method using data from ACTG175. Concluding remarks are given the in Section 5.

2.2 Estimation

Let $T^{\dagger}$ be a continuous failure time, $\mathbf{X}$ be a $p \times 1$ vector of bounded baseline predictors, and C be the censoring variable. For convenience, we assume that $\mathbf{X} =$

$(1, X_2, ..., X_p)^\intercal$. Let $\{(T_i^\dagger, \mathbf{X}_i, C_i), i = 1, ...,\text{n}\}$ be n independent copies of $(T^\dagger, \mathbf{X}, C)$. Due to censoring, for the i^{th} subject, we only observe $\mathscr{D} = (T_i, \mathbf{X}_i, \delta_i)$, where $T_i = min(T_i^\dagger, C_i), \delta_i = I(T_i^\dagger \leq C_i)$ and $I(\cdot)$ is the indicator function. We consider the setting where $T_i^\dagger$ is subject to heavy censoring, and auxiliary post-treatment q-dimensional variables $\mathbf{S}$, correlated with $T^\dagger$ to a certain degree, is taken beyond baseline at time $t_s < \tau$. Note that $\mathbf{S}$ is censored for those with $C_i < t_s$, and can be deemed as censored for subjects with $T_i^\dagger < t_s$ since the survival outcome has been observed before time t_s for these subjects. For convenience, we assume that C is independent of $\{T^\dagger, \mathbf{X}, \mathbf{S}\}$. In the situation where C might depend on $\mathbf{X}$ and $\mathbf{S}$, weights need to be constructed differently, but the proposed method would be implemented similarly.

2.2.1 Estimation procedure

To develop risk prediction rules for τ-year survival $Y_\tau = I(T^\dagger \leq \tau)$, we fit a time-specific generalized linear model

$$Pr(T^\dagger \leq \tau|\mathbf{X}) = Pr(Y_\tau = 1|\mathbf{X}) = g(\boldsymbol{\beta}^\intercal \mathbf{X}) \tag{2.1}$$

where $g(\cdot)$ is a known, strictly increasing, differentiable function and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters. Note that $\boldsymbol{\beta}$ is a function of τ and can be denoted as $\boldsymbol{\beta}_\tau$, but for convenience we keep the simple notation throughout. Under (2.1), one can use estimating equation $\mathbf{U}_0(\boldsymbol{\beta}) = E[\mathbf{X}(Y_\tau - g(\boldsymbol{\beta}^\intercal \mathbf{X}))] = 0$ to solve for $\boldsymbol{\beta}$ and we denote the solution to $\mathbf{U}_0(\boldsymbol{\beta})$ as $\bar{\boldsymbol{\beta}}$. With censoring, Uno et al. (2007) proposed to estimate $\bar{\boldsymbol{\beta}}$ using $\widetilde{\boldsymbol{\beta}}$, based on the inverse probability weighted (IPW) estimating function:

$$\tilde{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{\tau i} \mathbf{X}_i \{Y_{\tau i} - g(\boldsymbol{\beta}^T \mathbf{X}_i)\} \tag{2.2}$$

where $\hat{w}_{\tau i} = \frac{I(T_i \leq \tau)\delta_i + I(T_i > \tau)}{\hat{G}(T_i \wedge \tau)}$, and $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$. It was shown that $\tilde{\boldsymbol{\beta}}$ is a consistent estimator of $\bar{\boldsymbol{\beta}}$ even if the working model (2.1) is

incorrectly specified.

With such construction, $\hat{w}_{\tau i}$ is zero for the subjects censored before τ, and so the model estimation of $\widetilde{\boldsymbol{\beta}}$ could be quite inefficient in the heavy censoring setting. To improve the efficiency, in addition to include baseline information from these subjects, one could potentially incorporate the post-baseline covariate $\mathbf{S}$. The difficulty of incorporating $\mathbf{S}$ lies in the non-negligible censoring before t_s, naively ignoring which will cause bias. But note that $Y_\tau = I(T^\dagger \leq t_s) + I(t_s < T^\dagger \leq \tau)$, and therefore we propose to impute Y_τ such that $I(T^\dagger \leq t_s)$ and $I(t_s < T^\dagger \leq \tau)$ are estimated separately, where the second part is estimated only using subjects who are observed beyond t_s and thus incorporating $\mathbf{S}$ becomes feasible. Let $Y_{t_s} = I(T^\dagger \leq t_s)$ and $\mathbf{Z}^\intercal = (\mathbf{X}^\intercal, \mathbf{S}^\intercal)$. We suggest that $\mathbf{Z}$ is rescaled so that the diagonal entries of its covariance matrix are 1's. We take $\Phi(\mathbf{X})^T = \{1, \phi_1(X_2)^T, ..., \phi_p(X_p)^T\}$ where $\phi_k(X^k)$ be the set of basis for the k^{th} covariate in $\mathbf{X}$, and $\Phi(\mathbf{Z})^T = \{1, \phi_1(X_2)^T, ..., \phi_p(X_p)^T, \phi_{p+1}(S_1)^T, ..., \phi_{p+q}(S_q)^T\}$.

The proposed $\widehat{\boldsymbol{\beta}}$ is based on the estimating function

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\{g(\hat{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{\hat{G}(t_s)} g(\hat{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\boldsymbol{\beta}^\intercal \mathbf{X}_i)\} \tag{2.3}$$

where $\widehat{\boldsymbol{\theta}}_{t_s}$ is the minimizer of:

$$\widehat{\mathbf{Q}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{n}^{i=1} \hat{w}_{t_s i} \ell\left\{Y_{t_s i}, g(\boldsymbol{\theta}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\right\} + \lambda_{t_s n} \mathcal{Q}(|\boldsymbol{\theta}_{[-1]}|), \quad \text{with } 0 \leq \lambda_{t_s} = o(n^{-\frac{1}{2}}) \tag{2.4}$$

where $\hat{w}_{t_s i} = \frac{I(T_i \leq t_s)\delta_i + I(T_i > t_s)}{\hat{G}(T_i \wedge t_s)}$, $\ell(y, x) = y \log\{g(x)\} + (1-y)\log\{1-g(x)\}$, and $\mathcal{Q}(\cdot)$ is a penalty function such as the ridge or LASSO (Friedman et al., 2001). The amount of regularization is controlled by the tuning parameter λ_{t_s} that is chosen via the cross validation but restricted to the order of $o(n^{-\frac{1}{2}})$ to ensure the desired convergence rate.

Similarly, $\widehat{\boldsymbol{\gamma}}_\tau$ is the minimizer of

$$\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t_s)\frac{\hat{w}_{\tau i}}{\hat{S}(t_s)} \ell\left\{Y_{\tau i}, g(\boldsymbol{\gamma}^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\right\} + \lambda_\tau \mathcal{Q}(|\boldsymbol{\gamma}_{[-1]}|) \tag{2.5}$$

where $\hat{S}(t_s) = \frac{n_{t_s}}{n\hat{G}(t_s)}$ and $\hat{w}_{\tau i} = \frac{I(T_i \leq \tau)\delta_i + I(T_i > \tau)}{\hat{G}(T_i \wedge \tau)}$,
or equivalently

$$\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) = \frac{1}{n_{t_s}} \sum_{\Omega_{t_s}} \widehat{w}_{\tau i}^{\Omega_{t_s}} \ell \left\{ Y_{\tau i}, g(\boldsymbol{\gamma}^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) \right\} + \lambda_\tau \mathcal{Q}(|\boldsymbol{\gamma}_{[-1]}|)$$

where $\Omega_{t_s} = \{i : T_i > t_s\}$ and $\widehat{w}_{\tau i}^{\Omega_{t_s}}$ is estimated similarly as $\widehat{w}_{\tau i}$ but only within Ω_{t_s}. Note that even when there is no intermediate covariate available, using above procedure but only incorporating $\mathbf{X}$, by replacing $\Phi(\mathbf{Z}_i)$ with $\Phi(\mathbf{X}_i)$ in (2.3) and (2.5), might lead to higher efficiency, especially in the settings where the association between the baseline predictors and the risk varies by time. Appendix B.1 shows that $\widehat{\boldsymbol{\beta}}$ is a consistent estimate of $\bar{\boldsymbol{\beta}}$ and Appendix B.2 shows that $\widehat{\boldsymbol{\beta}}$ follows a normal distribution asymptotically.

The proposed method allows to incorporate S collected at a pre-specified time t_s to improve the efficiency. However, for any given intermediate covariates, it is unclear whether it would be predictive of the outcome and whether incorporating it would improve efficiency or cause overfitting. More importantly, if there are multiple time points when intermediate covariates are collected, it is unclear incorporating which of them would lead to most substantial efficiency gain. In order to avoid post-hoc decisions, we propose a systematic way to seek for the optimal linear combination of these estimators. Suppose there are m candidate consistent estimators $\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, ..., \widehat{\boldsymbol{\beta}}^{(m)}$ of $\bar{\boldsymbol{\beta}}$, each being $1 \times p$ vector, which could include $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}'s$ that incorporate intermediate covariates at different time points. For simplicity, we only consider component-wise linear combination. For each predictor X_j, $j = 1, ..., p$, among all the linear combinations $\beta_j^{comb} = (1 - \mathbf{1}^\intercal \mathbf{W}_j)\widehat{\beta}_j^{(1)} + \mathbf{W}_j^\intercal (\widehat{\beta}_j^{(2)}, ..., \widehat{\beta}_j^{(m)})^\intercal$, we pursue $\bar{\beta}_j^{comb}$ such that

$$\overline{\mathbf{W}}_j = \underset{\mathbf{W}_j}{\text{argmin}} Var(\beta_j^{comb}).$$

Note that $\beta_j^{comb} = \widehat{\beta}_j^{(1)} - \mathbf{W}_j^\intercal \widehat{\boldsymbol{\Delta}}_j$, where $\widehat{\boldsymbol{\Delta}}_j = (\widehat{\beta}_j^{(1)} - \widehat{\beta}_j^{(2)}, ..., \widehat{\beta}_j^{(1)} - \widehat{\beta}_j^{(m)})^\intercal$, and therefore

it can be viewed as a standard least square problem such that

$$\overline{\mathbf{W}}_j = \underset{\mathbf{W}_j}{\text{argmin}} E(\widehat{\beta}_j^{(1)} - \alpha_j - \mathbf{W}_j^{\intercal}\widehat{\mathbf{\Delta}}_j)^2,$$

where $\alpha_j = E(\widehat{\beta}_j^{(1)})$ is a nuisance parameter. Practically, one could generate a large number, say B, of resampling copies, $\widehat{\boldsymbol{\beta}}^{(1*)}, \widehat{\boldsymbol{\beta}}^{(2*)}, ..., \widehat{\boldsymbol{\beta}}^{(m*)}$, for each of $\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, ..., \widehat{\boldsymbol{\beta}}^{(m)}$, then

$$\widehat{\mathbf{W}}_j = \underset{\mathbf{W}_j}{\text{argmin}} \sum_{i=1}^{B} (\widehat{\beta}_{ji}^{(1*)} - \alpha_j - \mathbf{W}_j^{\intercal}\widehat{\mathbf{\Delta}}_{ji}^{*})^2,$$

where α_j is a nuisance parameter. When m is not small, one could consider the regularized estimation:

$$\widehat{\mathbf{W}}_j = \underset{\mathbf{W}_j}{\text{argmin}} \sum_{i=1}^{B} (\widehat{\beta}_{ji}^{(1*)} - \alpha_j - \mathbf{W}_j^{\intercal}\widehat{\mathbf{\Delta}}_{ji}^{*})^2 + \upsilon_j \mathcal{Q}(|\mathbf{W}_j|),$$

where υ_j is the tuning parameter to control the amount of regularization in the order of $o(B^{-\frac{1}{2}})$.

2.2.2 Inference via Resampling

Therefore we propose a perturbed resampling procedure to obtain the optimal combined estimator and its variance. Specifically, let $\mathbf{V} = (V_1, ..., V_n)^{\intercal}$ be an $n \times 1$ vector of independent and identically distributed random variables with mean 1 and variance 1, generate independent of $\mathscr{D}$. Then we obtain $\widehat{\boldsymbol{\beta}}^*$, a perturbed version of $\widehat{\boldsymbol{\beta}}$, based on the estimating function

$$\widehat{\mathbf{U}}_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} V_i \mathbf{X}_i \{ g(\widehat{\boldsymbol{\theta}}_{t_s}^{*T}\Phi(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{\hat{G}(t_s)} g(\hat{\boldsymbol{\gamma}}_{\tau}^{*T}\Phi(\mathbf{Z}_i)) - g(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i) \},$$

where $\widehat{\boldsymbol{\theta}}_{t_s}^*$ is the minimizer of

$$\widehat{\mathbf{Q}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{n}^{i=1} \widehat{w}_{t_s i}^* \ell \{ Y_{t_s i}, g(\boldsymbol{\theta}^{\intercal}\mathbf{\Phi}(\mathbf{X}_i)) \} + \lambda_{t_s} \mathcal{Q}(|\boldsymbol{\theta}_{[-1]}|), \quad \text{with } 0 \leq \lambda_{t_s} = o(n^{-\frac{1}{2}})$$

and $\hat{\boldsymbol{\gamma}}^*$ is the minimizer of

$$\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t_s) \frac{\widehat{w}_{\tau i}^*}{\hat{S}(t_s)} \ell \{ Y_{\tau i}, g(\boldsymbol{\gamma}^{\intercal}\mathbf{\Phi}(\mathbf{Z}_i)) \} + \lambda_{\tau} \mathcal{Q}(|\boldsymbol{\gamma}_{[-1]}|)$$

where $\widehat{w}_{t_s i} = \frac{I(T_i \leq t_s)\delta_i + I(T_i > t_s)}{\hat{G}(T_i \wedge t_s)}$.

In practice, one can generate B random samples of V to obtain B realizations of the candidate consistent estimates $\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, ..., \widehat{\boldsymbol{\beta}}^{(m)}$, then use the procedure described in the section 1.2 to construct their optimal linear combination $\widehat{\boldsymbol{\beta}}^{comb}$ and estimate the confidence interval. Note that the regular bootstrap sampling is similar to the above proposed perturbation procedure except that the weights are generated from multinomial distribution. Such approach could encounter instability issues in the rare event setting because the subsamples might have too few events. By assigning positive weights to all subject, the proposed perturbation procedure can avoid such problem.

2.3 Simulation

Simulation studies were conducted using 500 Monte Carlo datasets, each with sample size 500, to evaluate the performance of the proposed estimator and compare to existing methods. For each dataset, we use 500 perturbations (or bootstrap subsamples for other methods) to estimate the variance. For the proposed estimator, natural spline bases with pre-specified 3 knots for each covariate are used for the imputation model. In the section 2.3.1, we consider the scenario where the number of baseline predictors is small relative to the number of events, and in the section 2.3.2, we consider the case where additional unimportant baseline predictors are included and the regularization is demanded.

For comparison, we obtained (i) IPW estimator with censoring weights constructed using KM estimator (IPW_{KM}) as suggested in Uno et al. (2007) (ii) IPW estimator with censoring weights constructed from a Cox model including $\mathbf{X}$ only (IPW_{cox}) (iii) AIPW estimator constructed as in DiRienzo (2009) with censoring weights estimated by KM and use only $\mathbf{X}$ for the imputation model ($\text{AIPW}_{\text{KM},\mathbf{X}}$); (iv) AIPW estimator similar to (iv) but additionally incorporating $\mathbf{S}$ to the imputation

model by adding an indicator variable for each of $\mathbf{S}$ for missing values as handled in Lu and Tsiatis (2008) ($\text{AIPW}_{\text{KM},\mathbf{Z}}$). Note that there is no existing methods that can be directly implemented to incorporate intermediate covariates with non-negligible censoring in baseline prediction model estimation, and therefore $\text{AIPW}_{\text{KM},\mathbf{Z}}$ is constructed to show the results of naively handling the censored values. We assume $\mathbf{S}$, a 1×4 vector, is a single intermediate covariate collected at 4 different time points. S_j was constructed as having increasing correlation with the outcome over time, but also has increasing proportion of missing values due to censoring. Therefore it is unclear incorporating which S_j would achieve the highest efficiency. It is also possible that incorporating S_j might result in overfitting, and the proposed estimating procedure using only $\mathbf{X}$ but imputing the outcome piecewise would lead to higher efficiency gain. Therefore, for each timepoint j, we constructed the estimator using $\mathbf{X}$ only ($\text{AUG}^j_{\text{KM},\mathbf{X}}$) and the one using $\mathbf{Z}$ ($\text{AUG}^j_{\text{KM},\mathbf{Z}}$), respectively. We then obtained linear combination of IPW_{KM}, $AUG^j_{\text{KM},\mathbf{X}}$ and $AUG^j_{\text{KM},\mathbf{Z}}$ at all 4 time points ($\text{AUG}^{\text{CMB}}_{\text{KM}}$). To evaluate the incremental value of incorporating $\mathbf{S}$, in the low dimension setting, we also obtained optimal linear combination of $\text{AUG}^j_{\text{KM},\mathbf{X}}$ alone at all 4 time points and IPW_{KM} ($\text{AUG}^{\text{CMB}}_{\text{KM},\mathbf{X}}$).

2.3.1 Low dimension baseline model

We consider $p = 4$ and generated $\mathbf{X}_{-1}$, $T^\dagger$ and S from

$$\mathbf{X}_{-1} = (X_2, X_3, X_4)^\intercal \sim N(\mathbf{0}, 0.3 + 0.7\mathbb{I}_3), \quad \text{where } \mathbb{I}_d \text{ is the } d \times d \text{ diagonal matrix,}$$
$$\log(T^\dagger) = 0.5(X_2 + X_3 + X_4) + 0.5X_2^2 + X_3^2 + 0.5X_4^2 - 3 + \text{logit}(U) + \log(\alpha),$$
$$\text{with } U \sim \text{Uniform(0,1)},$$

Suppose the interest is in estimating the risk probability at $\tau = 0.8$. We considered (i) a low event rate (12−18% by τ) and heavy censoring setting (65−74% before τ) with $\{\alpha = 12, \lambda = 0.5\}$; and (ii) a moderate event rate (25−34% by τ) and moderate censoring setting (37−50% before τ) with $\{\alpha = 6, \lambda = 1\}$. We assume that a single intermediate covariate is collected at four different time points

$(t_1 = 0.05, t_2 = 0.1, t_3 = 0.15$ and $t_4 = 0.2)$, denoted as $S = \{S_1, S_2, S_3, S_4\}$. We generate S_j as

$$S_j = logit(U) + 0.1(X_1 + X_2) + \varepsilon/(10 * t_j^{1.5}), \quad \text{with } U \sim \text{Uniform(0,1) and } \varepsilon \sim N(0, 0.5),$$

such that the Kendall's correlation coefficient between S_j and Y_τ is about 13% for $j = 1$, 34% for $j = 2$, 55% for $j = 3$, and 65% for $j = 4$. The proportions of subjects who are observed beyond the four time points are approximately 87%, 76%, 67%, and 60%, respectively.

As shown in Table 2.1, the proposed estimator offers consistent estimation and gains substantial efficiency relative to IPW_{KM}. Despite efficiency loss due to incorrect imputation model, $AIPW_{\text{KM},\mathbf{X}}$ still provides consistent estimation due to the double robustness. However, when additionally incorporating $\mathbf{S}$ into the imputation model, $AIPW_{\text{KM},\mathbf{Z}}$ becomes a biased estimator due to the fact that the missingness of $\mathbf{S}$ depends on C. We note that $\text{AUG}^{\text{CMB}}_{\text{KM},\mathbf{X}}$ also has considerable efficiency gain due to expanded base functions and possibly more accurate piecewise imputation, but $\text{AUG}^{\text{CMB}}_{\text{KM}}$ further gains efficiency by additionally incorporating $\mathbf{S}$. Figure 2.1 shows that the proposed estimator at different timepoint has different efficiency gain, but the final estimator $\text{AUG}^{\text{CMB}}_{\text{KM}}$, the optimal combination of them, has the most efficiency gain. The resampling procedure works well with coverage percentage for 95% confidence interval range from 93% - 97%.

2.3.2 Baseline model regularization

The purpose of this simulation is to assess the performance of the proposed method in finite sample when p is not small relative to the number of events and regularization is used for feature selection. We consider $p = 11$ and adaptive LASSO (Zou, 2006; Zhang and Lu, 2007) for baseline model selection. $\mathbf{X}_{-1}$ is generated from $N(\mathbf{0}, \mathbb{I}_{10})$ and

$$\log(T^\dagger) = X_2 + X_3 + X_4 + 0.5X_2^2 + X_3^2 + 0.5X_4^2 - 3 + \text{logit}(U) + \log(6).$$

Figure 2.1: The percent of efficiency gain (%EffG) and the coverage percentage for 95% confidence interval for the proposed estimator in the low dimension baseline model

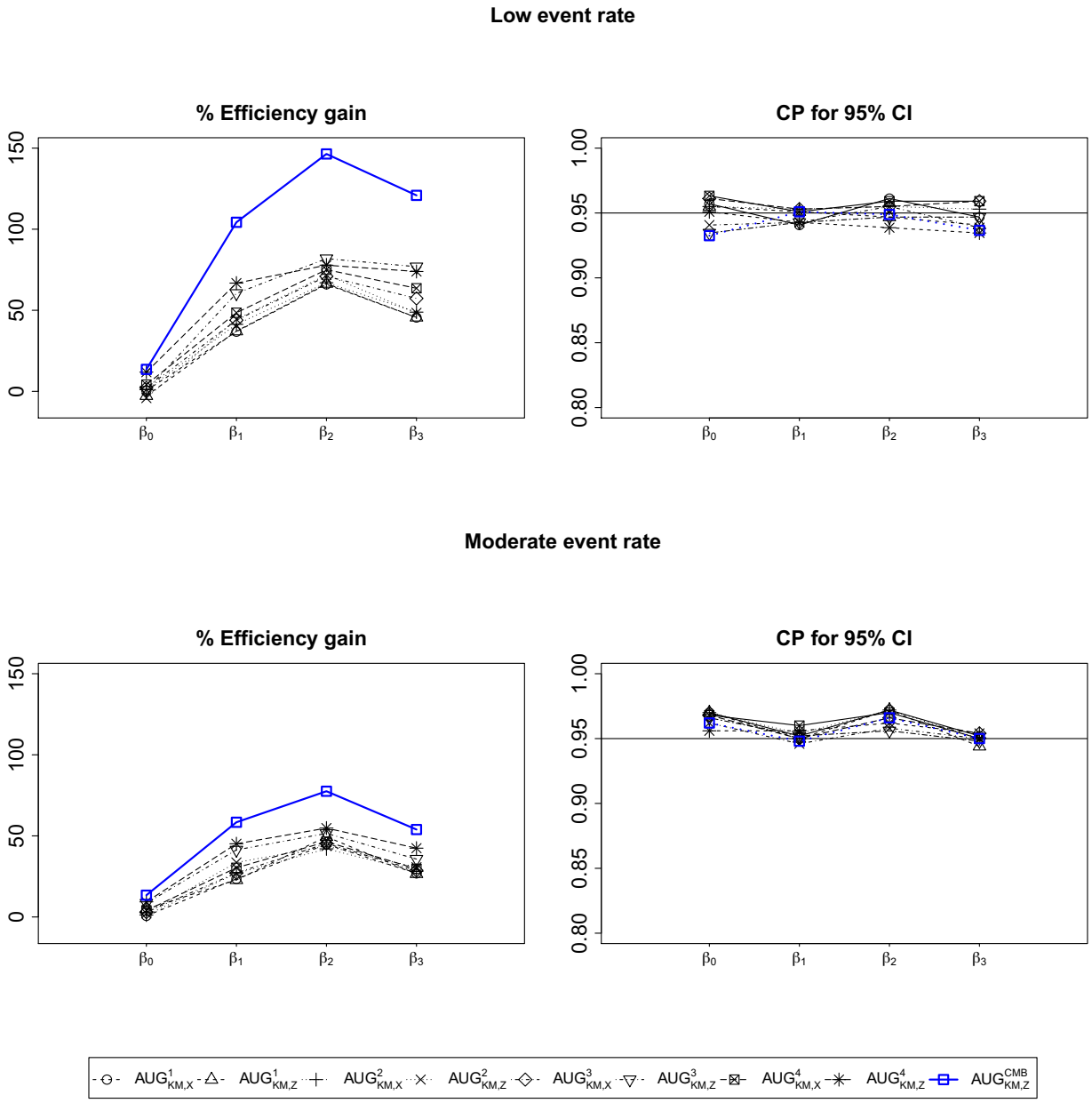


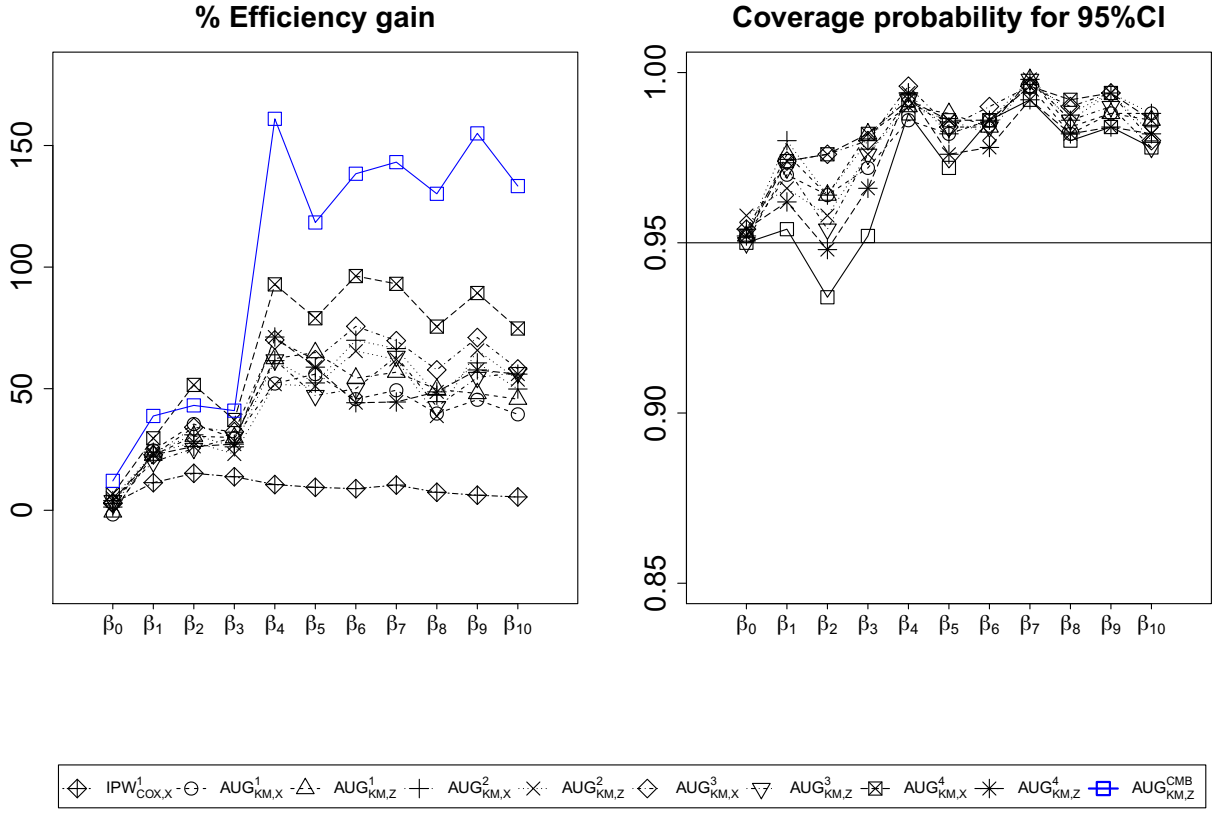
Table 2.1: Empirical bias, SE (ESE) and average of the estimated SE (ASE) for the low-dimensional setting. Shown also are the percent of efficiency gain (%EffG) relative to the IPW_{KM} estimator.

	Bias × 100				ESE × 100				%EffG			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Low Event Rate												
IPW_{KM}	1.22	-2.55	-1.20	-1.85	15.20	20.29	20.45	20.52	0.00	0.00	0.00	
IPW_{cox}	-0.37	-2.28	-1.58	-1.63	15.13	18.96	18.32	18.78	1.43	14.63	24.00	19.41
$\text{AIPW}_{\text{KM},\mathbf{X}}$	0.53	-3.50	-0.68	-1.86	15.74	28.14	29.94	28.65	-6.23	-48.01	-53.25	-48.50
$\text{AIPW}_{\text{KM},\mathbf{Z}}$	105.10	12.42	7.57	12.12	11.10	14.09	14.40	14.23	-97.92	18.47	58.46	21.54
$\text{AUG}^{\text{CMB}}_{\text{KM},\mathbf{X}}$	0.36	-0.66	-2.73	-0.42	14.90	16.15	14.62	15.42	5.84	60.93	90.98	78.34
$\text{AUG}^{\text{CMB}}_{\text{KM}}$	2.35	0.11	-2.83	0.25	14.18	14.32	12.79	13.86	12.47	103.79	144.39	120.76
Moderate Event Rate												
IPW_{KM}	0.68	-0.81	-0.71	-0.13	10.88	14.20	14.00	14.20	-	-	-	-
IPW_{cox}	0.03	-0.51	-0.71	0.00	10.79	13.46	12.98	13.42	2.02	11.58	16.29	12.01
$\text{AIPW}_{\text{KM},\mathbf{X}}$	0.75	-0.88	-0.64	-0.15	10.93	15.59	15.63	15.53	-1.13	-17.03	-19.72	-16.37
$\text{AIPW}_{\text{KM},\mathbf{Z}}$	46.68	6.93	3.75	6.75	9.86	12.03	12.32	11.77	-94.78	4.99	18.48	9.49
$\text{AUG}^{\text{CMB}}_{\text{KM},\mathbf{X}}$	-0.01	-0.18	-1.49	0.56	10.66	12.13	11.21	11.94	4.31	38.72	53.37	40.97
$\text{AUG}^{\text{CMB}}_{\text{KM}}$	0.24	0.79	-0.72	0.76	10.23	11.26	10.48	11.41	13.55	58.80	77.85	54.31

We generated C and $\mathbf{S}$ as in Section 2.3.1 and let $\lambda = 1$ and $\tau = 0.8$, leading to an observed event rate by τ of about $26 - 38\%$, such that the effective sample size is around 100-200, not large relative to $p = 11$. To control overfitting in the imputation model, L_1 penalized logistic regression (Friedman et al., 2010) with IPW was employed when estimating the imputation model.

Figure 2.2 summarizes the results for $AUG^j_{\text{KM},\mathbf{X}}$ and $AUG^j_{\text{KM},\mathbf{Z}}$ at each of the 4 time points, and the final combined linear optimal estimator $AUG^{\text{CMB}}_{\text{KM}}$ as well as IPW_{KM} and IPW_{cox} as benchmarks for efficiency assessment. The AIPW methods were not included as no associated regularization procedures were available. In general, $AUG^j_{\text{KM},\mathbf{Z}}$ is more efficient than IPW_{cox}, and the optimal linear combination of all time points has lead to the highest efficiency gain. The resampling procedures also perform well with empirical coverage percentage ranging from 92-95% for informative signals. The coverage percentage for the zero signals range from 96%-98%, which is expected owing to the oracle properties.

Figure 2.2: The percent of efficiency gain (%EffG) and the coverage percentage for 95% confidence interval for the proposed estimator for regularized baseline model



2.4 Example

We illustrate the proposed procedures using a dataset from the AIDS Clinical Trial Group (ACTG) Protocol 175 (Hammer et al., 1996). This study consists of 2467 patients randomized to 4 different treatments; zidovudine (ZDV) only, ZDV+didanosine (ZDV+DDI), ZDV+zalcitabine (ZDV+ZAL), and didanosine (DDI) only. Suppose the interest is to predict the risk of death or AIDS defining events by week 144, by when 38% subjects were censored and another 12% had an event observed. We consider the model with three indicator variables for treatment arms as well as the following baseline covariates: baseline CD4 counts (CD4), age(< 40 vs. > 40), the karnofsky score (ks), and symptomatic status (SS). For imputation model, we also include grade 3 and above toxicity and tolerability at week 48, 72, and 96 as intermediate covariates. Subjects who were still at risk at the three time points are $90\%, 85\%$, and 80%, respectively.

We scale all the covariates to have standard deviation 1. For the enriched imputation modeling required by our approach, we use spline bases with 3 knots for all continuous variables. Resampling with 500 replications is used to generate the variance of the IPW_{KM} method and our proposed methods, and bootstrap was used for other methods. As shown in Table 2.2, the point estimates from IPW_{KM} and IPW_{cox} are quite similar, indicating that censoring might be independent of the baseline predictors. The proposed estimator incorporating intermediate covariates at different time points and the final optimal linear combined estimator also provide comparable point estimation. As in the simulation, AIPW estimator that incorporate intermediate covariates but naively ignore the missing values ($AIPW_{\text{KM},\mathbf{Z}}$) leads to biased estimation which is quite different from others. The proposed estimator provides estimates similar to IPW_{KM} and IPW_{cox}, but have substantial smaller standard error. Again, $AIPW^{j}_{\text{KM},\mathbf{X}}$ and $AIPW^{j}_{\text{KM},\mathbf{Z}}$ have different degree of efficiency gain, but the combined linear optimal estimator has gained the most efficiency.

Table 2.2: Estimated intercept (Int) and covariate effects for the week 144 progression risk prediction models among those treated with ZVD and with ZVD+DDI arms from ACTG175.

	Coefficient$_{SE}$							
	Int	arm0	arm1	arm2	Age	CD4	Karnof	SS
IPW_{KM}	$-2.02_{.080}$	$0.20_{.085}$	$-0.08_{.091}$	$0.03_{.084}$	$-0.16_{.063}$	$-0.53_{.086}$	$-0.24_{.064}$	$0.25_{.062}$
IPW_{cox}	$-2.02_{.082}$	$0.19_{.085}$	$-0.08_{.091}$	$0.02_{.084}$	$-0.19_{.063}$	$-0.52_{.086}$	$-0.23_{.062}$	$0.23_{.061}$
$\text{AIPW}_{\text{KM},\mathbf{X}}$	$-2.02_{.081}$	$0.21_{.090}$	$-0.08_{.092}$	$0.03_{.089}$	$-0.14_{.066}$	$-0.52_{.085}$	$-0.25_{.068}$	$0.25_{.067}$
$\text{AIPW}_{\text{KM},\mathbf{Z}}$	$-1.08_{.046}$	$0.16_{.059}$	$-0.01_{.069}$	$0.05_{.061}$	$0.11_{.049}$	$-0.29_{.053}$	$-0.16_{.049}$	$0.16_{.048}$
$\text{AUG}^{48}_{\text{KM},\mathbf{X}}$	$-2.00_{.079}$	$0.17_{.073}$	$-0.09_{.073}$	$0.01_{.068}$	$-0.19_{.060}$	$-0.50_{.086}$	$-0.22_{.058}$	$0.23_{.057}$
$\text{AUG}^{48}_{\text{KM},\mathbf{Z}}$	$-1.99_{.079}$	$0.16_{.071}$	$-0.09_{.071}$	$-0.01_{.066}$	$-0.19_{.060}$	$-0.49_{.087}$	$-0.22_{.056}$	$0.22_{.056}$
$\text{AUG}^{72}_{\text{KM},\mathbf{X}}$	$-2.01_{.078}$	$0.18_{.069}$	$-0.08_{.067}$	$0.02_{.062}$	$-0.20_{.058}$	$-0.51_{.084}$	$-0.23_{.053}$	$0.24_{.055}$
$\text{AUG}^{72}_{\text{KM},\mathbf{Z}}$	$-2.00_{.077}$	$0.18_{.066}$	$-0.08_{.065}$	$0.02_{.059}$	$-0.20_{.056}$	$-0.51_{.082}$	$-0.23_{.053}$	$0.24_{.053}$
$\text{AUG}^{96}_{\text{KM},\mathbf{X}}$	$-2.01_{.078}$	$0.18_{.066}$	$-0.08_{.061}$	$0.02_{.059}$	$-0.20_{.055}$	$-0.51_{.084}$	$-0.23_{.052}$	$0.24_{.053}$
$\text{AUG}^{96}_{\text{KM},\mathbf{Z}}$	$-2.00_{.077}$	$0.17_{.064}$	$-0.08_{.059}$	$0.02_{.057}$	$-0.20_{.054}$	$-0.51_{.083}$	$-0.23_{.051}$	$0.24_{.052}$
$\text{AUG}^{\text{CMB}}_{\text{KM}}$	$-2.00_{.076}$	$0.15_{.056}$	$-0.08_{.044}$	$0.01_{.045}$	$-0.23_{.051}$	$-0.51_{.080}$	$-0.22_{.046}$	$0.23_{.048}$

2.5 Remarks

When the interest is to estimate a $\tau-$year risk prediction model given baseline covariates but the efficiency is lost due to heavy censoring, we propose a procedure to leverage the information from censored subjects and intermediate covariates to improve the estimation efficiency for such a model. This method is related to the one proposed by Zheng and Cai (2017) in the sense that they both use imputation augmentation to incorporate post-baseline covariates to improve efficiency for a τ-GLM baseline model, but the novelty of this method include: (1) it allows the incorporation of the intermediate covariates that have non-negligible missing values due to censoring or early failures and still provide consistent estimation for baseline prediction model; (2) when there are multiple time points when the intermediate covariates are collected, we propose a systematic way to obtain optimal linear combination of estimators incorporating these covariates at different timepoints.

We make the assumption that the censoring is dependent of the covariates for theoretical justification. When the censoring depends on the covariates, similar to other IPW-based approaches, the consistency of our estimation relies on the weight model that correctly capture the relationship between censoring and the covariates. Zheng and Cai (2017) has examined the settings where the censoring depends on the covariates and the weight model is mis-specified, and showed that the imputation-based estimator is more robust to weight mis-specificaiton than a simple IPW approach. Compared to other survival model such as Cox proportional hazard model, although it does not require the specification of the weight model, it has stronger model assumption and the resulting estimator actually depends on the censoring, the proposed estimator has the advantage of being censoring free and more robust to model mis-specification.

The efficiency gain of the proposed estimator relies on three parts: correction of the mis-specified working model, censored subjects before τ, and the intermedi-

ate covariates that provides additional information other than baseline predictors. Therefore, the method would gain efficiency in the settings with heavy censoring and strong prognostic intermediate factors.

Chapter 3

Deriving Optimal Individualized Treatment Rules from Randomized Studies for Maximizing T-year Survival Probability

Abstract

Individualized treatment rules are usually determined based on the baseline covariates with the goal to optimize population average outcome if the rules were followed in the entire population. When the outcome of interest is survival probability at a pre-specified time point t, most of existing methods identify optimal treatment rules by directly maximizing the population average outcome which could suffer numerical instability, especially when the censoring rate is high. We propose an inverse probability weighted non-parametric estimator for the value function based on an imposed working model, and an additional imputation based augmentation procedure that could significantly increase the efficiency of the estimator. The method guards against the possible model mis-specification as long as the predicted treatment difference from the imposed model is a monotone function of the true difference, and it also allows incorporating of post-baseline covariates to further improve efficiency. Numerical studies show that the proposed method performs better than existing methods in terms of numerical stability and higher efficiency, especially in heavy censoring settings. We also apply the method to ACTG175, a randomized clinical trial of HIV-infected subjects.

3.1 Introduction

Individualized medicine has received increasing attention recently. The results from randomized clinical trials comparing mean treatment difference do not necessarily provide evidence for an optimal individualized treatment rule (ITR). The treatment giving the better average outcome might not necessarily lead to better outcome at the individual level since subjects might respond differently to different treatments due to individual heterogeneity. When comparing two treatments, the ITR is a binary decision rule, typically a function of baseline covariates, that assigns subjects to one of the treatments based on their baseline information. An optimal ITR would optimize the population average outcome, also known as the value function, if followed for all patients.

A wide range of methods have been proposed to derive an ITR with a single baseline predictor (Song and Pepe, 2004; Bonetti and Gelber, 2004) or with multiple baseline predictors (Cai et al., 2011; Zhang et al., 2012a,2012b; Zhao et al., 2012; Matsouaka et al., 2014; Tian et al., 2014). Most of existing methods are for continuous or binary outcomes, while only a few consider survival outcomes with censored data. Zhao et al. (2015) proposed a doubly robust estimator for mean survival time by recasting the problem as a weighted misclassification rate. Jiang et al. (2016) extended the method by Zhang (2012b) to the survival setting where the problem is casted in a missing value framework and the t-year survival probability is estimated by an inverse probability weighted (IPW) Kaplan-Meier (KM) estimator. Bai et al. (2013, 2016) proposed a locally efficient augmented IPW (AIPW) estimator for the survival setting where the interest of outcome is a function of survival time. Even though kernel smoothing and weighted learning were introduced for these methods to ease computational difficulty, they are in a framework where directly maximizing the empirical value function could still be computationally prohibitive and numerically

unstable, especially in the setting of heavy censoring.

When the outcome of the interest is the survival status by a pre-specified time τ, Uno et al. (2007) proposed a time specific baseline prediction model (τ-GLM) that can be consistently estimated using an IPW framework. Zheng and Cai (2017) also proposed an augmented procedure to further improve the efficiency of estimation of such model in the heavy censoring setting. An ITR could then be developed based on the predicted risk level for subjects. Under possible model mis-specification, although τ-GLM can still be consistently estimated in the sense that the estimation converges to a determined censoring free parameter, solely relying on the predicted outcome might not lead to an optimal ITR that optimizes the population value function. However, the stability of τ-GLM could be useful. It might be reasonable to make the assumption that the imposed working model is not severely mis-specified such that the predicted risk difference is a monotone function of the true difference. This motivates us to propose a semi-nonparametric estimating procedure to estimate the optimal population value function and its corresponding ITR based on the predicted treatment difference from imposed τ-GLM models. The method could also easily incorporate post-baseline covariates to potentially further improve efficiency of estimation.

The rest of the paper is organized as follows: in Section 2, we describe the proposed estimation procedure including a numerical consideration for computation convenience; in Section 3, we present results from simulation studies showing that the proposed method has better numerical stability and efficiency than existing methods; in Section 4, we illustrate our method in an AIDS clinical trial ACTG175 that evaluated treatments for HIV-infected patients. Some concluding remarks are given in Section 5.

3.2 Estimation

Consider two study treatments $A = 0, 1$ available for randomization at baseline. Let $T^\dagger$ be a continuous failure time, $\mathbf{X}$ be a $p\times 1$ vector of bounded baseline predictors, and C be the censoring variable. For convenience, we assume that $\mathbf{X} = (1, X_2, ..., X_p)^\intercal$. Let $\{(A_i, T_i^\dagger, \mathbf{X}_i, C_i), i = 1, ..., n\}$ be n independent copies of $(A, T^\dagger, \mathbf{X}, C)$, with n_a subjects randomized to treatment $A = a$ where n_a/n goes to a constant in the open interval $(0, 1)$ as $n \to \infty$. Due to censoring, for the i^{th} subject, we only observe $\mathscr{D} = (A_i, T_i, \mathbf{X}_i, \delta_i)$, where $T_i = min(T_i^\dagger, C_i), \delta_i = I(T_i^\dagger \leq C_i)$ and $I(\cdot)$ is the indicator function. Suppose auxiliary post-treatment q-dimensional variable $\mathbf{S}$, correlated with $T^\dagger$, is taken beyond baseline but early during follow-up so that it is available for everyone. For convenience, we assume that C is independent of $\{T^\dagger, \mathbf{X}, \mathbf{S}\}$ conditioning on A. The assumption can be relaxed such that C could depend on $\{\mathbf{X}, \mathbf{S}\}$, and the proposed method could be implemented similarly but requires an estimated relation between C and $\{\mathbf{X}, \mathbf{S}\}$.

3.2.1 Estimation procedure

Define the binary τ-year survival outcome $Y = I(T^\dagger \leq \tau)$. To develop treatment specific risk prediction rules for τ-year survival outcome, we fit a time-specific generalized linear model

$$Pr(T^\dagger \leq \tau | \mathbf{X}, A = a) = Pr(Y = 1 | \mathbf{X}, A = a) = g(\boldsymbol{\beta}_a^\intercal \mathbf{X}) \tag{3.1}$$

where $a = 0, 1$, $g(\cdot)$ is a known, strictly increasing, differentiable function and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters (recall that $\mathbf{X}$ is a p-dimensional covariate including 1 as the first element). Note that both Y and $\boldsymbol{\beta}$ is τ-specific and can be denoted as Y_τ and $\bar{\boldsymbol{\beta}}_\tau$, respectively, but for convenience we keep the simple notation throughout. Under (3.1), one can use the following estimating equation: $\overline{\mathbf{U}}_a(\boldsymbol{\beta}) = E[\mathbf{X}(Y - g(\boldsymbol{\beta}^\intercal \mathbf{X}))|A = a] = 0$ to solve for $\boldsymbol{\beta}_a$ and we denote the solution to

$\overline{\mathbf{U}}_a(\boldsymbol{\beta})$ as $\bar{\boldsymbol{\beta}}_a$. With censoring, one could estimate $\bar{\boldsymbol{\beta}}_a$ using $\widehat{\boldsymbol{\beta}}_a$, based on the inverse probability weighted (IPW) estimating function (Uno et al., 2007):

$$\widehat{\mathbf{U}}_a(\boldsymbol{\beta}) = \frac{1}{n_a}\sum_{i=1}^{n} I(A_i = a)\widehat{w}_{ai}\mathbf{X}_i\{Y_i - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\} \tag{3.2}$$

where $\hat{w}_{ai} = \frac{I(T_i \le \tau)\delta_i + I(T_i > \tau)}{\widehat{G}_a(T_i \wedge \tau)}$, and $\widehat{G}_a(\cdot)$ is the Kaplan-Meier estimator of $G_a(\cdot)$. Uno et al. (2007) has shown that $\widehat{\boldsymbol{\beta}}_a$ is a consistent estimator of $\bar{\boldsymbol{\beta}}_a$ even if the working model (3.1) is incorrectly specified. The model-based treatment difference can then be obtained as $\widehat{D}_{\mathbf{X}} = g(\widehat{\boldsymbol{\beta}}_1^{\mathsf{T}}\mathbf{X}) - g(\widehat{\boldsymbol{\beta}}_0^{\mathsf{T}}\mathbf{X})$, which is a consistent estimator of $D_{\mathbf{X}} = g(\bar{\boldsymbol{\beta}}_1^{\mathsf{T}}\mathbf{X}) - g(\bar{\boldsymbol{\beta}}_0^{\mathsf{T}}\mathbf{X})$.

Define $Y^{(a)}$ as the potential outcome if a patient receives treatment a. $Y^{(a)}$ is known as the counterfactual outcome, and typically not observed for $Y^{(1-A)}$, the treatment that a subject is not assigned to. Let $\mathcal{I}$ be a binary ITR such that $\mathcal{I} = 1$ assigns patients to treatment 1, and $\mathcal{I} = 0$ assigns patients to treatment 0. Define $Y^{(\mathcal{I})} = Y^{(1)}\mathcal{I} + Y^{(0)}(1-\mathcal{I})$, the τ-year risk if a patient receives treatment assignment according to $\mathcal{I}$. An optimal ITR would maximize the population value function with respect to the survival probability

$$\mathcal{O}(\mathcal{I}) = E\{1 - Y^{(\mathcal{I})}\} = E\{(1 - Y^{(1)})\mathcal{I} + (1 - Y^{(0)})(1-\mathcal{I})\}$$

Note that it is equivalent to minimize the population value function with respect to the risk $E\{Y^{(1)}\mathcal{I} + Y^{(0)}(1-\mathcal{I})\}$. When $\mathcal{I}$ is a function of $\mathbf{X}$, it is not difficult to see that the ITR that optimizes $\mathcal{O}(\mathcal{I})$ is the Bayes rule

$$\underset{\mathcal{I}}{\text{argmax}}\mathcal{O}(\mathcal{I}) = \mathcal{I}_{\Delta_{\mathbf{X}}}(0) = I(\Delta_{\mathbf{X}} > 0)$$

where $\Delta_{\mathbf{X}} = E\{Y^{(1)} - Y^{(0)}|\mathbf{X}\}$. If the working model 3.1 is correctly specified for both treatments, then $D_{\mathbf{X}} = \Delta_{\mathbf{X}}$ and $\mathcal{I}_{\widehat{D}_{\mathbf{X}}}(0) = I(\widehat{D}_{\mathbf{X}} > 0)$ can be directly used as the optimal ITR. The assumption that the working model is correctly specified is often unrealistic. However, it might not be unreasonable to assume that $D_{\mathbf{X}}$ is a monotone function of $\Delta_{\mathbf{X}}$ if the working model is not severely misspecified. With

that assumption, for any cutoff point δ on $\Delta_{\mathbf{X}}$, there exists a cutoff point d on $D_{\mathbf{X}}$ such that

$$\mathcal{O}(\mathcal{I}_{D_{\mathbf{X}}}(d)) = \mathcal{O}(\mathcal{I}_{\Delta_{\mathbf{X}}}(\delta))$$

and the optimal value function is

$$\mathcal{O}(\mathcal{I}_{\Delta_{\mathbf{X}}}(0)) = \underset{d}{\text{argmax}}\mathcal{O}(\mathcal{I}_{D_{\mathbf{X}}}(d))$$

Practically, one could estimate $\hat{\mathcal{O}}(\mathcal{I}_{\widehat{D}_{\mathbf{X}}}(d))$ for all $d \in \{\widehat{D}_{\mathbf{X}}\}$ and identify $\widehat{d^{opt}} = \underset{d}{\text{argmax}}\hat{\mathcal{O}}(\mathcal{I}_{\widehat{D}_{\mathbf{X}}}(d))$.

Let $\hat{\mathcal{O}}_j = \hat{\mathcal{O}}(\mathcal{I}_{\widehat{D}_{\mathbf{X}}}(\widehat{D}_{\mathbf{X}j})), j = 1, ..., n$ and the maximum value function $\hat{\mathcal{S}} = \max \hat{\mathcal{O}}_j$. Suppose there was no censoring and all patients' counterfactual outcomes were observed, one could directly estimate $\mathcal{O}_j$ as

$$\hat{\mathcal{O}}_j^0 = \frac{\sum_{i=1}^n \mathbb{I}(\widehat{D}_{\mathbf{X}i} \leq \widehat{D}_{\mathbf{X}j})(1 - Y_i^{(1)}) + \sum_{i=1}^n \mathbb{I}(\widehat{D}_{\mathbf{X}i} > \widehat{D}_{\mathbf{X}j})(1 - Y_i^{(0)})}{n},$$

and $\hat{\mathcal{S}}^0 = \max \hat{\mathcal{O}}_j^0$. Typically the counterfactual outcome cannot be observed and the censoring is almost unavoidable in the studies with the survival outcome. One could then use the IPW estimator

$$\hat{\mathcal{O}}_j^{\text{IPW}} = \frac{\sum_{i=1}^n I(\widehat{D}_{\mathbf{X}i} \leq \widehat{D}_{\mathbf{X}j}) A_i \widehat{w}_{1i}(1 - Y_i)}{n_1} + \frac{\sum_{i=1}^n I(\widehat{D}_{\mathbf{X}i} > \widehat{D}_{\mathbf{X}j})(1 - A_i)\widehat{w}_{i0}(1 - Y_i)}{n_0},$$

and $\hat{\mathcal{S}}^{\text{IPW}} = \max \hat{\mathcal{O}}_j^{\text{IPW}}$. This estimator could be further improved in terms of efficiency by incorporating the baseline covariates.

Specifically, we propose to estimate $\mathcal{O}_j$ as

$$\hat{\mathcal{O}}_j^{\text{AUG}} = \frac{\sum_{i=1}^n \mathbb{I}(\widehat{D}_{\mathbf{X}i} \leq \widehat{D}_{\mathbf{X}j})(1 - \widehat{Y}_i^{(1)}) + \sum_{i=1}^n \mathbb{I}(\widehat{D}_{\mathbf{X}i} > \widehat{D}_{\mathbf{X}j})(1 - \widehat{Y}_i^{(0)})}{n}$$

where $\widehat{Y}_i^{(a)}$ is imputed semi-parametrically by incorporating baseline covariates and $\hat{\mathcal{S}}^{\text{AUG}} = \max \hat{\mathcal{O}}_j^{\text{AUG}}$. Specifically,

$$\widehat{Y}_i^{(a)} = \widehat{Y}_i^{(a)}(\mathbf{X}) = g(\widehat{v}^{(a)}(\widehat{D}_{\mathbf{X}i}) + \mathbf{\Phi}(\mathbf{X}_i)^{\intercal}\hat{\gamma}^{(a)}) \tag{3.3}$$

where $\mathbf{\Phi}(\mathbf{X}_i)$ is a finite set of basis function for $\mathbf{X}_i$ and $\{\widehat{v}^{(a)}(\widehat{D}_{\mathbf{X}i}), \widehat{\boldsymbol{\gamma}}^{(a)}\}$ simultaneously satisfy:

$$\widehat{\boldsymbol{\gamma}}^{(a)} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \widehat{w}_{ai} \ell_{bin}(Y_i, g(\widehat{v}^{(a)}(\widehat{D}_{\mathbf{X}i}) + \mathbf{\Phi}(\mathbf{X}_i)^{\intercal}\boldsymbol{\gamma})) + \lambda_n \mathcal{Q}(|\boldsymbol{\gamma}_{[-1]}|)$$

$$\sum_{i=1}^{n=i} I(A_i = a) K_h(\widehat{D}_{\mathbf{X}i} - d) \widehat{w}_{ai} \{Y_i - g(v^{(a)}(d) + \mathbf{\Phi}(\mathbf{X}_i)^{\intercal}\widehat{\boldsymbol{\gamma}}^{(a)})\} = 0 \quad \forall\, d \in \{\widehat{D}_{\mathbf{X}i}\}$$

where $\ell_{bin}(y, g(x)) = y \log\{g(x)\} + (1-y)\log\{1-g(x)\}$, $\mathcal{Q}(\cdot)$ is a penalty function such as the ridge or LASSO, and $0 \leq \lambda_n = o(n^{-\frac{1}{2}})$. The first estimating function imposes a working imputation model, and the second equation uses a non-parametric kernel method to ensure that $\widehat{Y}_i^{(a)}$ is a consistent estimator for $E[Y^{(a)}|\widehat{D}_{\mathbf{X}}]$ regardless of the adequacy of the imposed model. One would need to iteratively solve the above two equations to obtain $\{\widehat{v}^{(a)}(\widehat{D}_{\mathbf{X}i}), \widehat{\boldsymbol{\gamma}}^{(a)}\}$, which could be numerically challenging. It can be shown that (3.3) is numerically equivalent to $\widehat{Y}^{(a)}(\mathbf{X})_i = g\{\mathbf{H}(\widehat{D}_{\mathbf{X}i}, \mathbf{X}_i)^{\intercal}\hat{\boldsymbol{\alpha}}^{(a)}\}$ and $\hat{\boldsymbol{\alpha}}^{(a)}$ is the minimizer of

$$\sum_{i=1}^{n} \widehat{w}_i \ell_{bin}(Y_i, g(\mathbf{H}(\widehat{D}_{\mathbf{X}i}, \mathbf{X}_i)^{\intercal}\boldsymbol{\alpha}^{(a)}) + \lambda_n \mathcal{Q}(|\boldsymbol{\alpha}_{[-1]}|)$$

where $\mathbf{H}(\widehat{D}_{\mathbf{X}i}, \mathbf{X}_i) = (\mathbf{B}(\widehat{D}_{\mathbf{X}i})^{\intercal}, \mathbf{\Phi}(\mathbf{X}_i)^{\intercal})^{\intercal}$ and $\mathbf{B}(\widehat{D}_{\mathbf{X}i})$ is the spline basis for $\widehat{D}_{\mathbf{X}i}$ with the number of the knots in the order of $n^{\frac{1}{5}}$. Such a numerical adaptation to a fully parametric model greatly reduces the computation cost of the proposed estimator.

3.2.2 Incorporating post-baseline covariates

When there are post-baseline intermediate covariates $\mathbf{S}$ available, one could potentially incorporate $\mathbf{S}$ to further improve the efficiency. First, the efficiency of $\widehat{D}_{\mathbf{X}}$ can be improved using the augmentation method proposed by Zheng and Cai (2017). Second, caution needs to be taken when incorporating $\mathbf{S}$ in the imputation models in 3.2.1 because $\mathbf{S}$ is collected post-randomization and could depend on A. Therefore, $\mathbf{S}$ cannot be used to impute the counterfactual outcome, but can be used to impute

the outcome that the patient is randomized to. Specifically, let $\mathbf{Z} = \{\mathbf{X}, \mathbf{S}\}$,

$$\widehat{Y}^{(A_i)}(\mathbf{Z}_i) = g(\widehat{v}^{(A_i)}(\widehat{D}_{\mathbf{X}i}) + \mathbf{\Phi}(\mathbf{Z}_i)^\intercal \hat{\gamma}^{(a)})$$

and with the numerical adaption with splines bases for the non-parametric function $\widehat{v}^{(a)}(\widehat{D}_{\mathbf{X}i})$, it is equivalent to:

$$\widehat{Y}^{(A_i)}(\mathbf{Z})_i = g(\mathbf{H}(D_{\mathbf{X}i}, \mathbf{Z}_i)^\intercal \hat{\boldsymbol{\alpha}}^{(A_i)}).$$

Therefore, the imputation of $Y_i^{(a)}$ is

$$\widehat{Y}_i^{(a)} = \mathcal{I}(A_i = a)\widehat{Y}^{(A_i)}(\mathbf{Z}_i)_i + \mathcal{I}(A_i = 1 - a)\widehat{Y}^{(a)}(\mathbf{X}_i).$$

Essentially, one could impute the counterfactual outcome by incorporating $\mathbf{X}$, the baseline covariates only, and impute the outcome for the treatment that the patient is randomized to by incorporating $\mathbf{Z}$, both baseline and post-baseline covariates.

3.3 Simulation

Numerical studies were carried out to evaluate the performance of the proposed estimators, each including 500 Monte-Carlo replications. Within each replication, a dataset of sample size 10000 is generated, and is divided as training set (N=500) and the test set (N=9500). For the training set, we consider randomized treatment assignment where $P(A = 1) = 0.5$. The ITR is evaluated using the training set (based on within-sample optimized value function), and then is applied to the test set to examine the out-of-sample performance of the ITR in terms of value function.

We generate the covariate $\mathbf{X}_{-1} = (X_2, X_3, X_4)^\intercal$, where X_2 follows a uniform distribution on [-2, 2] and $(X_3, X_4)^\intercal$ is a mean 0 multivariate normal distribution with standard deviation 1 and correlation coefficient 0.3. The hazard function is generated by $\lambda(t|X, A) = e^t \times exp\{0.5 + X_2 - X_3 - 0.5X_4^2 - A \times (1.5 + 3X_2 - 2X_3 + X_4^2)\}$ and C

is from $Uniform(0, C_c)$. Let $\tau = 1$. We tried settings where C_c=8, 2, and 1.5, corresponding to censoring rates of $20\%, 55\%$, and 80%. S is generated from $N(0, 0.5)$ for $T_i^\dagger < \tau$ and $N(1.5, 0.5)$ for $T_i^\dagger > \tau$. The wrong model includes $\{1, X_3, X_4\}$, missing an important covariate X_2 and quadratic term for X_4. Note that with such simulation setting, the underlying true optimal value function (maximum survival probability) at τ=1 is 0.85.

We obtain the estimation for the value function using (1) AIPW estimator as proposed by Bai et al. (2016) ($\hat{\mathcal{S}}^{\text{Avs}}$); (2) AIPW estimator as proposed by Jiang et al. (2016) ($\hat{\mathcal{S}}^{\text{KMvs}}$); (3) $\hat{\mathcal{S}}^{\text{IPW}}$ as described in 1.2; (4) $\hat{\mathcal{S}}^{\text{AUG}}$ as described in 1.2 with the imputation only include baseline covariates $\mathbf{X}$ ($\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$) (5) $\hat{\mathcal{S}}^{\text{AUG}}$ as described in 3.2.2 that incorporate post-baseline covariate $\mathbf{S}$ in addition to $\mathbf{X}$ ($\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{Z})$). To impute the counterfactual outcome, a 3-knots natural spline basis function is used to construct $\Phi(\mathbf{X})$ and $\Phi(\mathbf{Z})$, and a 8-knots natural spline basis function is used for $\mathbf{B}(D_{\mathbf{X}i})$. For each estimator, we use the training set to estimate within-sample and out-of-sample optimal value function, with the former examining the estimation for the optimal value function, and the latter assessing the performance of the ITR. The results are shown in Table 3.1

All methods perform similarly in the setting of low censoring rate. $\hat{\mathcal{S}}^{\text{Avs}}$ does not perform well in either estimating the optimal value function or ITR when the censoring rate increases. $\hat{\mathcal{S}}^{\text{KMvs}}$ estimator also tends to have slightly larger bias in the optimal value function estimation when the censoring rate increases, although the standard error is comparable to the proposed estimator. In addition, the out-of-sample estimation of the optimal value function is less efficient than the proposed method, suggesting that the ITR derived from the proposed method is better than the one derived from $\hat{\mathcal{S}}^{\text{KMvs}}$. For the proposed methods, $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$ is much more efficient than $\hat{\mathcal{S}}^{\text{IPW}}$ by imputing counterfactual outcome. When the censoring rate is low, $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$ that incorporates additional $\mathbf{S}$ does not lead to much additional gain

Table 3.1: Within sample and out-of-sample value function estimation and standard error

	%Censor	Correct model		Wrong model	
		within sample	out-of-sample	within sample	out-of-sample
$\hat{\mathcal{S}}^{\text{Avs}}$	20	$0.84_{.0276}$	$0.84_{.0183}$	$0.70_{.0309}$	$0.69_{.0180}$
$\hat{\mathcal{S}}^{\text{KMvs}}$		$0.86_{.0216}$	$0.84_{.0074}$	$0.72_{.0286}$	$0.70_{.0103}$
$\hat{\mathcal{S}}^{\text{IPW}}$		$0.86_{.0425}$	$0.84_{.0085}$	$0.72_{.0383}$	$0.70_{.0120}$
$\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$		$0.85_{.0212}$	$0.85_{.0042}$	$0.71_{.0275}$	$0.70_{.0068}$
$\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{Z})$		$0.85_{.0209}$	$0.85_{.0042}$	$0.71_{.0271}$	$0.70_{.0069}$
$\hat{\mathcal{S}}^{\text{Avs}}$	55	$0.68_{.0610}$	$0.76_{.0656}$	$0.60_{.0509}$	$0.65_{.0441}$
$\hat{\mathcal{S}}^{\text{KMvs}}$		$0.86_{.0246}$	$0.84_{.0092}$	$0.73_{.0305}$	$0.70_{.0113}$
$\hat{\mathcal{S}}^{\text{IPW}}$		$0.86_{.0491}$	$0.84_{.0104}$	$0.72_{.0449}$	$0.69_{.0154}$
$\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$		$0.84_{.0252}$	$0.85_{.0045}$	$0.70_{.0311}$	$0.70_{.0077}$
$\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{Z})$		$0.84_{.0229}$	$0.85_{.0044}$	$0.70_{.0287}$	$0.70_{.0074}$
$\hat{\mathcal{S}}^{\text{Avs}}$	80	$0.68_{.0652}$	$0.76_{.0689}$	$0.59_{.0557}$	$0.65_{.0473}$
$\hat{\mathcal{S}}^{\text{KMvs}}$		$0.87_{.0268}$	$0.84_{.0100}$	$0.73_{.0328}$	$0.70_{.0125}$
$\hat{\mathcal{S}}^{\text{IPW}}$		$0.86_{.0532}$	$0.84_{.0112}$	$0.73_{.0532}$	$0.69_{.0179}$
$\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$		$0.84_{.0280}$	$0.85_{.0047}$	$0.70_{.0358}$	$0.70_{.0091}$
$\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{Z})$		$0.84_{.0245}$	$0.85_{.0045}$	$0.70_{.0311}$	$0.70_{.0084}$

compared to $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$; but when censoring rate increases, $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$ is more efficient. The empirical standard error of the proposed estimators is approximated well with Bootstrap resampling with coverage probability for 95% confidence interval range between 93%-97%.

3.4 Example

We illustrate the proposed procedures using a dataset from the AIDS Clinical Trial Group (ACTG) Protocol 175 (Hammer et al., 1996). This study consisted of 2467 patients randomized to 4 different treatments; zidovudine (ZDV) only, ZDV+didanosine (ZDV+DDI), ZDV+zalcitabine (ZDV+ZAL), and didanosine (DDI) only. Suppose the outcome of interest is death or AIDS and the interest is to derive the optimal ITR to maximize the event free probability. We illustrate the proposed method for deriving optimal ITR for ZDV only vs. ZDV+DDI as well as DDI only vs. ZDV+DDI. We consider $\tau = 144$ weeks, by when 38% subjects were censored. The Kaplan-Meier estimates for the survival probabilities are 0.8208, 0.8638, and 0.8868 for ZDV only,

ZDV+ZAL and ZDV+DDI, respectively. We consider the working model with the baseline CD4 count (CD4), age, and the Karnofsky score (ks). For imputation model, we also include a 50%+ CD4 reduction from baseline and grade 3 and above toxicity and tolerability at week 24. Note that less than 0.5% subjects experienced an event of interest and another 3.5% subjects were censored before week 24. These subjects were excluded from the analyses. All the continuous variables were scaled to have standard deviation 1.

Table 3.2: Regression coefficients from treatment specific τ-GLM for the construction of the index score

	Intercept	Age	CD4	Karnof
ZDV only	-2.14	0.16	-0.70	-0.45
ZDV+ZAL	-1.74	0.30	-0.66	-0.16
ZDV+DDI	-2.22	-0.00	-0.56	-0.30

Table 3.3: ACTG175 Treatment assignment and optimal population proportion without events by Week 144

		$\hat{\mathcal{S}}_{(SE)}$	Proportion assigned ZDV+ZAL$_{(SE)}$
ZDV only vs. ZDV+ZAL	IPW	$0.8832_{(0.0238)}$	$0.8001_{(0.2046)}$
	AUG($\mathbf{X}$)	$0.8758_{(0.0143)}$	$0.9243_{(0.1461)}$
	AUG($\mathbf{Z}$)	$0.8756_{(0.0138)}$	$0.9535_{(0.1382)}$
ZDV+DDI vs. ZDV+ZAL	IPW	$0.8909_{(0.0264)}$	$0.3813_{(0.3215)}$
	AUG($\mathbf{X}$)	$0.8912_{(0.0127)}$	$0.4307_{(0.3085)}$
	AUG($\mathbf{Z}$)	$0.8890_{(0.0125)}$	$0.3873_{(0.2963)}$

Table 3.2 shows the regression coefficients ($\widehat{\boldsymbol{\beta}}_a$) from treatment specific τ-GLMs that are used to construct the index scores. Age appears to have no effect for ZDV+DDI arm, and the Karnof scores have smaller effect size for ZDV+ZAL arm. The estimated difference in survival probabilities are the index scores used to identify the optimal ITRs. Table 3.3 shows that $\hat{\mathcal{S}}^{\text{IPW}}$, $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$, and $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{Z})$ provided similar estimates on the maximum survival probability for both groups. But $\hat{\mathcal{S}}^{\text{AUG}}(\mathbf{X})$ has a much smaller standard error by incorporating baseline covariates to the impu-

tation. Incorporating additional post-baseline covariates at week 24 does not lead to significant efficiency gain. This could be due to various reasons: 1. the chosen covariates might not be strongly associated with the outcome of interest; 2. even the intermediate covariates is associated with the outcome of the interest, they might not contribute to the value function estimation with additional information that is not in the baseline covariates. The optimal survival probability from the derived ITR is around 0.88 for ZDV only vs. ZDV+ZAL, with majority of subjects assigned to ZDV+ZAL arm. For ZDV+DDI vs. ZDV+ZAL, the optimal survival probability is around 0.89 and 37-48% of subjects are assigned ZDV+ZAL arm. Comparing them to the KM estimates of the survival probabilities (0.8208 in ZDV only arm, 0.8638 in ZDV+ZAL arm, and 0.8868 in ZDV+DDI arm), the optimal ITR only has modest improvement on the survival probability in this particular case. It suggests that individualized medicine might not be necessary in this particular example.

3.5 Remarks

We propose a method to estimate the optimal value function for survival probability and the corresponding optimal ITR in randomized clinical trials with the time to event type of endpoints. Semi-parametric imputation based augmentation is introduced to improve the efficiency of the estimator and a numerical equivalent parametric approach is considered for computational convenience. In the heavy censoring setting where post-randomization intermediate covariates might be strongly predictive of the outcome, the method also allows incorporation of these covariates to further improve the efficiency. However, whether there will be additional efficiency gain from the intermediate covariates depends on its association with the outcome of interest as well as the baseline covariate. Numerical studies also assess the performance of the derived ITR by implementing the ITR to the validation data and evaluate the population survival probability. The proposed method is shown to have more consistent

estimation in the optimal value function and better ITR in terms of efficiency.

Appendix A

Appendix A for Chapter 1

Notations

To ensure that $\mathbf{U}_0(\boldsymbol{\beta}) = 0$ has a unique solution, we assume that there does not exist a $\boldsymbol{\beta}$ such that $P(\boldsymbol{\beta}^\intercal \mathbf{X}_1 > \boldsymbol{\beta}^\intercal \mathbf{X}_2 | T_1^\dagger \leq t \leq T_2^\dagger) = 1$ as in Uno et al. (2007). In addition, we assume that $\mathbf{Z}$ is bounded and the conditional distribution of $T^\dagger$ given $\mathbf{Z}$ is continuously differentiable. We also require standard assumptions about the censoring distribution such that $G(\tau) > 0$, $\sup_{t \leq \tau} |\widehat{G}(t) - G(t)| \xrightarrow{\mathcal{P}} 0$, and $n^{\frac{1}{2}}\{\widehat{G}(t) - G(t)\} = n^{-\frac{1}{2}} \sum_{i=1}^n \psi_i(t) + o_p(1)$ which converges weakly to a zero-mean Gaussian process, where $\psi_i(t) = \int_0^t \pi(u)^{-1} dM_{ci}(u)$ with $\pi(u) = P(T_i > u)$, $M_{ci}(u) = I(T_i \leq u, \delta_i = 0) - \int_0^u I(T_i > v) d\Lambda_c(v)$ and $\Lambda_c(\cdot)$ is the cumulative hazard function for the censoring variable C. Throughout, we define $\mathbb{J} = E[\mathbf{X}_i^{\otimes 2} \dot{g}(\bar{\boldsymbol{\beta}}_\tau^\intercal \mathbf{X}_i)]$, $\mathbf{F}_{1i} = \mathbb{J}^{-1}\mathbf{X}_i(Y_{\tau i} - g(\bar{\boldsymbol{\beta}}_\tau^\intercal \mathbf{X}_i))$, $\mathbf{F}_{2i} = \mathbb{J}^{-1}\mathbf{X}_i(Y_{\tau i} - g(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}_i)$, and $\boldsymbol{\mu}_{F_k}(s) = E\{\mathbf{F}_{ki} I(T_i^\dagger > s)\}$ for $k = 1, 2$. For any matrix $\mathbf{A}$, $\mathbf{A}_j$, represents the $j$the jth row vector, and for any a vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\intercal$. We let $\xrightarrow{\mathcal{P}}$ and $\xrightarrow{\mathcal{D}}$ denote convergence in probability and in distribution, respectively.

A.1 Allowing Censoring to Depend on Covariates

When C is potentially dependent on $\mathbf{X}$, we may extend the proposed methods to account for such dependency by imposing a semi-parametric model for C given $\mathbf{X}$

similar to Lin et al. (2001). For example, one may assume a Cox model,

$$G_{\mathbf{X}}(t) \equiv P(C \geq t \mid \mathbf{X}) = \exp\{-\exp(-\boldsymbol{\alpha}_0^{\mathsf{T}}\mathbf{X})\Lambda_{0C}(t)\},$$

for the censoring distribution and subsequently the censoring weights $\mathcal{W}_i = \{I(T_i \leq \tau)\delta_i + I(T_i > \tau)\}/G_{\mathbf{X}_i}(T_i \wedge \tau)$ can be consistently estimated as $\widehat{\mathcal{W}}_i = \{I(T_i \leq \tau)\delta_i + I(T_i > \tau)\}/\widehat{G}_{\mathbf{X}_i}(T_i \wedge \tau)$. For inference, one may modify the perturbation procedures to perturb $\widehat{\mathcal{W}}_i$ as $\widehat{\mathcal{W}}_i^*$, where $\widehat{\mathcal{W}}_i = \{I(T_i \leq \tau)\delta_i + I(T_i > \tau)\}/\widehat{G}^*_{\mathbf{X}_i}(T_i \wedge \tau)$, $\widehat{G}^*_{\mathbf{X}_i}(t) = \exp\{-\exp(-\mathbf{X}^{\mathsf{T}}\widehat{\boldsymbol{\alpha}}^*)\widehat{\Lambda}^*_{0C}(t)\}$, $\widehat{\boldsymbol{\alpha}}^*$ and $\widehat{\Lambda}^*_{0C}$ may be obtained by either weighting the Cox model fitting using $\mathbf{V}$ or explicitly as those given in the Web Appendix of Cai et al. (2010). Then the proposed procedures detailed in the manuscript can be updated using these new censoring weights to account for covariate dependent censoring.

A.2 Consistency of $\hat{\boldsymbol{\beta}}_\tau$

To show consistency of $\widehat{\boldsymbol{\beta}}_\tau$ for $\bar{\boldsymbol{\beta}}_\tau$, it suffices to show that (i) $\sup_{\boldsymbol{\beta}} |\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_0^c(\boldsymbol{\beta})| = o_p(1)$; and (ii) $\bar{\boldsymbol{\beta}}_\tau$ is a unique solution to $\mathbf{U}_0^c(\boldsymbol{\beta}) = 0$ (Newey and McFadden, 1994), where

$$\mathbf{U}_0^c(\boldsymbol{\beta}) = E[\mathbf{X}_i\{g(\bar{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}_i) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\}].$$

To show (i), let $\bar{\boldsymbol{\theta}}_\tau$ be the unique solution to

$$\mathbf{R}_0(\boldsymbol{\theta}) = E[\boldsymbol{\Phi}_i\{Y_{\tau i} - g(\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\Phi}_i)\}]$$

and $\mathbf{U}_n^c(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{X}_i\{g(\bar{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}_i) - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)\}$. Since $\lambda_n \to 0$, it follows from similar arguments as those given in Uno et al. (2007) that $\widehat{\boldsymbol{\theta}}_\tau \xrightarrow{\mathcal{P}} \bar{\boldsymbol{\theta}}_\tau$. By a uniform law of larger numbers (Pollard, 1990), $\sup_{\boldsymbol{\beta}} |\mathbf{U}_n^c(\boldsymbol{\beta}) - \mathbf{U}_0^c(\boldsymbol{\beta})| \xrightarrow{\mathcal{P}} 0$. On the other hand $\sup_{\boldsymbol{\beta}} |\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_n^c(\boldsymbol{\beta})| = o_p(1)$ follows directly from the consistency of $\widehat{\boldsymbol{\theta}}_\tau$ for $\bar{\boldsymbol{\theta}}_\tau$. Thus, $\sup_{\boldsymbol{\beta}} |\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_0^c(\boldsymbol{\beta})| = o_p(1)$. For (ii), we note that $\partial\mathbf{U}_0^c(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathsf{T}}$ is positive definite and $\mathbf{U}_0^c(\bar{\boldsymbol{\beta}}_\tau) = \mathbf{U}_0^c(\bar{\boldsymbol{\beta}}_\tau) - \mathbf{U}_0(\bar{\boldsymbol{\beta}}_\tau) = E(\mathbf{X}[Y_\tau - g\{\bar{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z})\}])$. Because each X_j is a linear combination of $\boldsymbol{\Phi}(\mathbf{Z})$ and $\mathbf{R}_0(\bar{\boldsymbol{\theta}}_\tau) = E(\boldsymbol{\Phi}(\mathbf{Z})[Y_\tau - g\{\bar{\boldsymbol{\theta}}_\tau^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z})\}]) = 0$, we have $\mathbf{U}_0^c(\bar{\boldsymbol{\beta}}_\tau) = 0$ and thus (ii) holds. This implies the consistency of $\widehat{\boldsymbol{\beta}}_\tau$.

A.3 Asymptotic Distribution of $\hat{\boldsymbol{\beta}}_\tau$

From a Taylor expansion and Law of Large Numbers,

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} (\mathbf{F}_{1i} - \mathbf{F}_{2i}) + \mathbb{J}^{-1}\mathbf{A} n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_\tau - \bar{\boldsymbol{\theta}}_\tau) + o_p(1)$$

where $\mathbf{A} = E[\mathbf{X}\boldsymbol{\Phi}(\mathbf{Z})^\intercal \dot{g}\{\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z})\}]$. By a Taylor series expansion for $\widehat{\mathbf{R}}_n(\widehat{\boldsymbol{\theta}}_\tau)$ around $\bar{\boldsymbol{\theta}}_\tau$, the asymptotic expansion for $\widehat{G}(t)$ and the fact that $n^{\frac{1}{2}}\lambda_n \to 0$, we have

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_\tau - \bar{\boldsymbol{\theta}}_\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{B}^{-1}\Big(w_i \boldsymbol{\Phi}_i \{Y_{\tau i} - g(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}_i)\} + \int_0^\tau \psi_i(s) dE[\boldsymbol{\Phi}_i \{Y_{\tau i} - g(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}_i)\} I(T_i^\dagger \le s)]\Big) + o_p(1),$$

where $\mathbf{B} = E[\boldsymbol{\Phi}(\mathbf{Z})^{\otimes 2} \dot{g}\{\bar{\boldsymbol{\theta}}_\tau \boldsymbol{\Phi}(\mathbf{Z})\}]$. It follows that

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \Big(\mathbf{F}_{1i} - \mathbf{F}_{2i} + w_i \mathbb{J}^{-1}(\mathbf{A}\mathbf{B}^{-1}\boldsymbol{\Phi}_i)\{Y_{\tau i} - g(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}_i)\} + + \mathbb{J}^{-1} \int_0^\tau \psi_i(s) dE[(\mathbf{A}\mathbf{B}^{-1}\boldsymbol{\Phi}_i)\{Y_{\tau i} - g(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}_i)\} I(T_i^\dagger \le s)]\Big) + o_p(1).$$

To simplify the above expansion, we note that

$$\begin{aligned}
[\mathbf{A}\mathbf{B}^{-1}]_{j,} &= E\left[X_j \boldsymbol{\Phi}(\mathbf{Z})^\intercal \dot{g}\{\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z})\}\right] [E\{\boldsymbol{\Phi}(\mathbf{Z})^{\otimes 2} \dot{g}(\bar{\boldsymbol{\theta}}_\tau \boldsymbol{\Phi}(\mathbf{Z}))\}]^{-1} \\
&= \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, E[\dot{g}(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}))\{X_j - \boldsymbol{\alpha}^\intercal \boldsymbol{\Phi}(\mathbf{Z})\}^2], \qquad \text{for } j = 1, ..., p.
\end{aligned}$$

Since each X_j is a linear combination of $\boldsymbol{\Phi}(\mathbf{Z})$, $\min_{\boldsymbol{\alpha}} E[\dot{g}(\bar{\boldsymbol{\theta}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}))\{X_j - \boldsymbol{\alpha}^\intercal \boldsymbol{\Phi}(\mathbf{Z})\}^2] = 0$ and thus we have $\mathbf{X}_i = \mathbf{A}\mathbf{B}^{-1}\boldsymbol{\Phi}_i$. This implies that

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left\{ \mathbf{F}_{1i} + (w_i - 1)\mathbf{F}_{2i} - \int_0^\tau \psi_i(s) \boldsymbol{\mu}_{F_2}(ds) \right\} + o_p(1).$$

It then follows from a central limit theorem that $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\Sigma}_{\text{AUG}})$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\text{AUG}} &= E[\{\mathbf{F}_{1i} + (w_i - 1)\mathbf{F}_{2i} - \int_0^\tau \psi_i(s) \boldsymbol{\mu}_{F_2}(ds)\}^{\otimes 2}] \\
&= \text{var}(\mathbf{F}_{1i}) + \text{var}\{(w_i - 1)\mathbf{F}_{2i}\} + \int_0^\tau \{\boldsymbol{\mu}_{F_2}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_2}(s)^{\otimes 2}\} \frac{d\Lambda_c(s)}{\pi(s)}. \qquad \text{(A.1)}
\end{aligned}$$

To compare the variance forms of $\widehat{\boldsymbol{\beta}}_\tau$ and $\widetilde{\boldsymbol{\beta}}_\tau$, we note that from Uno et al. (2007),

$$n^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left\{ \mathbf{F}_{1i} + (w_i - 1)\mathbf{F}_{1i} - \int_0^\tau \psi_i(s) d\boldsymbol{\mu}_{F_1}(s) \right\} + o_p(1),$$

which converges in distribution to a zero mean multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{\text{IPW}} = \text{var}(\mathbf{F}_{1i}) + \text{var}\{(w_i - 1)\mathbf{F}_{1i}\} + \int_0^\tau \{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_1}(s)^{\otimes 2}\} \frac{d\Lambda_c(s)}{\pi(s)}. \quad \text{(A.2)}$$

Now taking the difference between (A.2) and (A.1), we have that:

$$\begin{aligned}\Delta\text{var} = \boldsymbol{\Sigma}_{\text{IPW}} - \boldsymbol{\Sigma}_{\text{AUG}} =& \text{var}\{(w_i - 1)\mathbf{F}_{1i}\} - \text{var}\{(w_i - 1)\mathbf{F}_{2i}\} + \\ & \int_0^\tau \left[\{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_1}(s)^{\otimes 2}\} - \{\boldsymbol{\mu}_{F_2}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_2}(s)^{\otimes 2}\} \right] \frac{d\Lambda_c(s)}{\pi(s)}.\end{aligned}$$

Note that

$$\begin{aligned}
&\text{var}\{(w_i - 1)\mathbf{F}_{1i}\} - \text{var}\{(w_i - 1)\mathbf{F}_{2i}\} \\
&= E[(w_i - 1)^2 \mathbf{F}_{1i}^{\otimes 2}] - E[(w_i - 1)^2 \mathbf{F}_{2i}^{\otimes 2}] \\
&= E[(w_i^2 - 1)(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})] \\
&= E([\{\frac{I(T_i^* \le \tau) I(T_i^* \le C_i)}{G(T_i^*)}\}^2 + \{\frac{I(T_i^* > \tau) I(C_i > \tau)}{G(\tau)}\}^2 - 1](\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})) \\
&= E[\{\frac{I(T_i^* \le \tau)}{G(T_i^*)} + \frac{I(T_i^* > \tau)}{G(\tau)} - 1\}(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})] \\
&= E[\{(\frac{1}{G(\tau)} - 1) - (\frac{1}{G(\tau)} - \frac{1}{G(T_i^*)}) I(T_i^* \le \tau)\}(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})] \\
&= E[\{\int_0^\tau d\frac{1}{G(s)} - \int_0^\tau I(T_i^* \le s) d\frac{1}{G(s)}\}(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})] \\
&= E[\{\int_0^\tau I(T_i^* > s) d\frac{1}{G(s)}\}(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})] \\
&= \int_0^\tau E(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2} \mid T_i^\dagger > s) S^2(s) \frac{d\Lambda_c(s)}{\pi(s)},
\end{aligned}$$

and $\boldsymbol{\mu}_{F_k}(s)^{\otimes 2} = E(\mathbf{F}_{ki} \mid T_i^\dagger > s)^{\otimes 2} S^2(s)$. Therefore,

$$\begin{aligned}\Delta\text{var} = & \int_0^\tau [Var\{\mathbf{F}_{1i} | T_i^\dagger > s\} - Var\{\mathbf{F}_{2i} | T_i^\dagger > s\}] \frac{S^2(s) d\Lambda_c(s)}{\pi(s)} \\ & + \{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_2}(\tau)^{\otimes 2}\} \int_0^\tau \frac{d\Lambda_c(s)}{\pi(s)}.\end{aligned}$$

A.4 Asymptotic Distribution of $\widetilde{\mathscr{B}}_\tau$ and $\widehat{\mathscr{B}}_\tau$ in Adaptive Lasso

Without loss of generality, let $\mathscr{A} = \{j : \bar{\beta}_{\tau j} \neq 0\} = \{1, 2, ..., p_0\}$ and $p_0 < p$. Then $\mathbb{J}$ can be written as $\mathbb{J} = \begin{bmatrix} \mathbb{J}_{11} & \mathbb{J}_{12} \\ \mathbb{J}_{21} & \mathbb{J}_{22} \end{bmatrix}$, where $\mathbb{J}_{11}$ is a $p_0 \times p_0$ matrix. Let $\mathbf{X}^{\mathscr{A}} = \{1, X_2, ..., X_{p_0}\}$ corresponds to the non-zero $\bar{\beta}'_{\tau j}s$. Let $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}_\tau + n^{-\frac{1}{2}}\mathbf{u}$ and $\phi(x) = \log(1 + e^x)$, then $\widetilde{L}_n(\mathbf{u}) = \sum_{i=1}^n \hat{w}_i\{-Y_{\tau i}\mathbf{X}_i^\intercal(\bar{\boldsymbol{\beta}}_\tau + n^{-\frac{1}{2}}\mathbf{u}) + \phi[\mathbf{X}_i^\intercal(\bar{\boldsymbol{\beta}}_\tau + n^{-\frac{1}{2}}\mathbf{u})]\}$. Define

$$\tilde{\Gamma}_n(\mathbf{u}) = \widetilde{L}_n(\mathbf{u}) + \widetilde{\nu}_n \sum_{j=2}^{p} |\bar{\beta}_{\tau j} + n^{-\frac{1}{2}}u_j|/|\tilde{\beta}_{\tau j}|.$$

Let $\widetilde{\mathbf{u}}_\tau = \sqrt{n}(\widetilde{\mathscr{B}}_\tau - \bar{\boldsymbol{\beta}}_\tau) = \text{argmin}_{\mathbf{u}}\, \tilde{\Gamma}_n(\mathbf{u}) = \text{argmin}_{\mathbf{u}}\, \{\tilde{\Gamma}_n(\mathbf{u}) - \tilde{\Gamma}_n(0)\}$, and by Taylor series expansion:

$$\tilde{\Gamma}_n(\mathbf{u}) - \tilde{\Gamma}_n(0) = \sum_{i=1}^{n} \hat{w}_i\{n^{-\frac{1}{2}}A_1^{(i)} + n^{-1}A_2^{(i)} + n^{-\frac{3}{2}}A_3^{(i)}\} + A_4^{(n)}$$

where $A_1^{(i)} = (-Y_{\tau i} + g(\mathbf{X}_i^T\bar{\boldsymbol{\beta}}))\mathbf{X}_i^T\mathbf{u}$, $A_2^{(i)} = \frac{1}{2}\dot{g}(\mathbf{X}_i^T\bar{\boldsymbol{\beta}})\mathbf{u}^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{u}$, $A_3^{(i)} = \frac{1}{6}\ddot{g}(\mathbf{X}_i^\intercal\bar{\boldsymbol{\beta}}_{\tau *})(\mathbf{X}_i^\intercal\mathbf{u})^3$, $|\bar{\boldsymbol{\beta}}_\tau - \bar{\boldsymbol{\beta}}_{\tau *}| \leq |n^{-\frac{1}{2}}\mathbf{u}|$, and $A_4^{(n)} = \lambda_n \sum_{j=2}^{p}\{|\bar{\beta}_{\tau j} + n^{-\frac{1}{2}}u_j| - |\bar{\beta}_{\tau j}|\}/|\hat{\beta}_{\tau j}|$.

Using Martingale representation,

$$\begin{aligned}
\tilde{\Gamma}_n(\mathbf{u}) - \tilde{\Gamma}_n(0) &\approx \int_0^t \frac{\sqrt{n}(G(s) - \hat{G}(s))}{G(s)} d\sum_{i=1}^{n} w_i(n^{-1}A_1^{(i)} + n^{-\frac{3}{2}}A_2^{(i)} + n^{-2}A_3^{(i)})I(T_i \leq s) \\
&\quad + \sum_{i=1}^{n} w_i\{n^{-\frac{1}{2}}A_1^{(i)} + n^{-1}A_2^{(i)} + n^{-\frac{3}{2}}A_3^{(i)}\} + A_4^{(n)} \\
&\approx -n^{-\frac{1}{2}}\sum_{i=1}^{n} \mathbf{u}^\intercal\mathbb{J}\boldsymbol{\eta}_{\text{IPW},i} + n^{-1}\sum_{i=1}^{n} w_i A_2^{(i)} + n^{-\frac{3}{2}}\sum_{i=1}^{n} w_i A_3^{(i)} + A_4^{(n)}
\end{aligned}$$

Using similar arguments in Zou (2006), the first term $\xrightarrow{\mathcal{D}} \mathbf{u}^\intercal\tilde{\mathbf{W}}$ with $\tilde{\mathbf{W}} \sim \mathbf{N}\{\mathbf{0}, \text{var}(\mathbb{J}\boldsymbol{\eta}_{\text{IPW},i})\}$, the second term $\xrightarrow{\mathcal{P}} \frac{1}{2}\mathbf{u}^\intercal\mathbb{J}\mathbf{u}$, the third term $\xrightarrow{\mathcal{P}} 0$, and the last term

$$A_4^{(n)} \xrightarrow{\mathcal{P}} \begin{cases} 0 & \text{if } \bar{\beta}_j \neq 0 \\ 0 & \text{if } \bar{\beta}_{\tau j} = 0 \text{ and } \mathbf{u}_j = 0 \\ \infty & \text{if } \bar{\beta}_{\tau j} = 0 \text{ and } \mathbf{u}_j \neq 0. \end{cases}$$

Thus, by Slutsky's theorem, we see that $\tilde{\Gamma}_n(\mathbf{u}) - \tilde{\Gamma}_n(0) \xrightarrow{\mathcal{D}} H(\mathbf{u})$ for every u, where

$$H(\mathbf{u}) = \begin{cases} \frac{1}{2}\mathbf{u}_{\mathscr{A}}^{\intercal}\mathbb{J}_{11}\mathbf{u}_{\mathscr{A}} - \mathbf{u}_{\mathscr{A}}^{\intercal}\tilde{\mathbf{W}}_{\mathscr{A}} & \text{if } u_j = 0 \ \forall j \notin \mathscr{A} \\ \infty & \text{otherwise.} \end{cases}$$

Because $\tilde{\Gamma}_n(\mathbf{u}) - \tilde{\Gamma}_n(0)$ is convex and the unique minimum of H is $(\mathbb{J}_{11}^{-1}\tilde{\mathbf{W}}_{\mathscr{A}}, \mathbf{0})^{\intercal}$, $\tilde{\mathbf{u}}_{\mathscr{A}} \xrightarrow{\mathcal{D}} \mathbb{J}_{11}^{-1}\tilde{\mathbf{W}}_{\mathscr{A}}$ and $\tilde{\mathbf{u}}_{\mathscr{A}^c} \xrightarrow{\mathcal{D}} \mathbf{0}$, where $\mathbb{J}_{11}^{-1}\tilde{\mathbf{W}}_{\mathscr{A}} = \mathbf{N}(\mathbf{0}, E(\boldsymbol{\eta}_{\text{IPW},i}^{\mathscr{A}\otimes 2}))$ with $\boldsymbol{\eta}_{\text{IPW},i}^{\mathscr{A}} = \mathbf{F}_{1i}^{\mathscr{A}} + (w_i - 1)\mathbf{F}_{1i}^{\mathscr{A}} + \int_0^t \psi_i(s) d\boldsymbol{\mu}_{\mathbf{F}_{1i}^{\mathscr{A}}}(s)$ and $\mathbf{F}_{1i}^{\mathscr{A}} = \mathbb{J}_{11}^{-1}\mathbf{X}_i^{\mathscr{A}}(Y_{\tau i} - g(\mathbf{X}_i^{\intercal}\bar{\boldsymbol{\beta}}_{\tau}))$.

For the proposed augmented estimator $\widehat{\mathscr{B}}_{\tau}$, note that $\widehat{\mathbf{u}} = \sqrt{n}(\widehat{\mathscr{B}}_{\tau} - \bar{\boldsymbol{\beta}}_{\tau}) = \text{argmin}_{\mathbf{u}} \hat{\Gamma}_n(\mathbf{u})$, where $\hat{\Gamma}_n(\mathbf{u}) = \widehat{L}_n(\mathbf{u}) + \widehat{\nu}_n \sum_{j=2}^p |\bar{\beta}_{\tau j} + n^{-\frac{1}{2}}u_j|/|\hat{\beta}_{\tau j}|$ and $\widehat{L}_n(\mathbf{u}) = \sum_{i=1}^n [-g(\boldsymbol{\Phi}_i^{\intercal}\hat{\boldsymbol{\theta}})\mathbf{X}_i^{\intercal}(\bar{\boldsymbol{\beta}}_{\tau} + n^{-\frac{1}{2}}\mathbf{u}) + \phi\{\mathbf{X}_i^{\intercal}(\bar{\boldsymbol{\beta}}_{\tau} + n^{-\frac{1}{2}}\mathbf{u})\}]$. It can be shown that

$$\hat{\Gamma}_n(\mathbf{u}) - \hat{\Gamma}_n(\mathbf{0}) \approx -n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{u}^{\intercal}\mathbb{J}\boldsymbol{\eta}_{\text{AUG},i} + n^{-1}\sum_{i=1}^n w_i A_2^{(i)} + n^{-\frac{3}{2}}\sum_{i=1}^n w_i A_3^{(i)} + A_4^{(n)}.$$

From similar arguments as for $\tilde{u}$, $\hat{\mathbf{u}}_{\mathscr{A}} \xrightarrow{\mathcal{D}} \mathbb{J}_{11}^{-1}\hat{\mathbf{W}}_{\mathscr{A}}$ and $\hat{\mathbf{u}}_{\mathscr{A}^c} \xrightarrow{\mathcal{D}} \mathbf{0}$, where $\mathbb{J}_{11}^{-1}\hat{\mathbf{W}}_{\mathscr{A}} \sim \mathbf{N}(\mathbf{0}, E(\boldsymbol{\eta}_{\text{AUG},i}^{\mathscr{A}\otimes 2})$ with $\boldsymbol{\eta}_{\text{AUG},i}^{\mathscr{A}} = \mathbf{F}_{1i}^{\mathscr{A}} + (w_i - 1)\mathbf{F}_{2i}^{\mathscr{A}} + \int_0^t \psi_i(s) d\boldsymbol{\mu}_{\mathbf{F}_{2i}^{\mathscr{A}}}(s)$ and $\mathbf{F}_{2i}^{\mathscr{A}} = \mathbb{J}_{11}^{-1}\mathbf{X}_i^{\mathscr{A}}\{Y_{\tau i} - g(\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\theta}}_{\tau})\}$.

A.5 Appendix Tables

Table A.1: Empirical bias, SE (ESE), average of the estimated SE (ASE),coverage probabilities (CovP) of the 95% CIs, and the percent of efficiency gain (%EffG) relative to the IPW_{cox} estimator for model ($\mathscr{M}_1$) with censoring distributions ($\mathscr{C}_{\text{cox}}$) or ($\mathscr{C}_{\text{ncox}}$).

Censoring		Bias × 100				ESE_{ASE}× 100				CovP				%EffG			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
	IPW_{cox}	-1.0	0.2	0.2	-0.2	$16.2_{17.8}$	$18.3_{18.3}$	$16.9_{16.6}$	$17.2_{17.6}$	97	97	97	97				
	IPW_{KM}	7.0	28.3	2.6	1.1	$17.0_{16.9}$	$18.3_{19.5}$	$18.1_{18.1}$	$18.2_{18.9}$	93	71	95	96	-21.9	-70.6	-15.1	-11.2
	AIPW_{KM}	7.2	21.2	3.3	2.7	$17.1_{14.7}$	$17.9_{14.8}$	$20.3_{13.5}$	$17.8_{14.2}$	87	68	83	89	-23.1	-56.6	-32.8	-8.0
$\mathscr{C}_{\text{cox}}$	AIPW_{cox}	1.3	2.8	0.1	0.3	$17.2_{14.8}$	$19.1_{14.8}$	$20.1_{13.7}$	$17.7_{14.4}$	91	87	83	89	-11.4	-10.0	-29.3	-5.1
	AUG_X	3.6	12.2	-2.8	-1.3	$16.8_{16.4}$	$17.7_{17.3}$	$14.4_{14.9}$	$15.5_{16.5}$	94	91	97	97	-10.6	-27.6	32.7	22.8
	$\text{AUG}_{X,S}$	3.9	5.6	-2.1	-0.3	$14.4_{13.5}$	$14.5_{13.4}$	$13.2_{12.4}$	$13.3_{12.9}$	93	93	95	96	19.0	39.2	58.9	69.1
	AUG_{CMB}	3.9	5.7	-2.0	-0.2	$14.3_{13.4}$	$15.2_{13.3}$	$13.2_{12.3}$	$13.2_{12.9}$	93	91	94	97	21.2	27.2	59.6	70.9
	$\text{AUG}_X^{\text{cox}}$	0.3	-0.7	-1.3	-0.3	$15.8_{16.9}$	$17.3_{16.1}$	$14.6_{14.4}$	$15.3_{15.9}$	96	94	96	97	5.5	11.7	32.0	27.3
	$\text{AUG}_{X,S}^{\text{cox}}$	2.7	0.9	-2.1	-0.5	$14.5_{13.7}$	$15.0_{13.2}$	$13.8_{12.3}$	$13.3_{13.0}$	94	92	93	96	21.2	49.0	46.5	68.0
	$\text{AUG}_{\text{CMB}}^{\text{cox}}$	2.6	0.1	-1.7	-0.4	$14.6_{13.6}$	$14.8_{13.0}$	$13.4_{12.1}$	$13.2_{12.9}$	93	92	93	96	20.4	52.3	55.9	70.8
	IPW_{cox}	-9.5	-8.8	2.1	1.0	$12.9_{13.8}$	$12.8_{11.5}$	$13.0_{12.6}$	$13.1_{13.4}$	92	84	95	96				
	IPW_{KM}	12.1	38.5	4.2	2.8	$13.7_{13.7}$	$13.3_{13.4}$	$13.0_{13.3}$	$13.0_{13.8}$	87	16	94	96	-22.8	-85.4	-6.9	-3.4
	AIPW_{KM}	7.6	26.8	6.5	4.4	$13.6_{12.3}$	$12.9_{11.9}$	$13.8_{11.0}$	$13.1_{11.7}$	88	39	85	92	6.6	-72.6	-25.7	-9.9
$\mathscr{C}_{\text{ncox}}$	AIPW_{cox}	-6.3	-3.6	2.5	1.3	$13.6_{12.4}$	$14.4_{11.7}$	$14.7_{11.0}$	$14.0_{11.8}$	90	89	87	92	15.1	9.4	-22.2	-13.9
	AUG_X	12.1	15.1	-3.1	-1.1	$13.3_{13.2}$	$12.7_{12.5}$	$11.2_{11.7}$	$11.9_{12.8}$	84	81	95	97	-20.1	-37.5	29.0	19.6
	$\text{AUG}_{X,S}$	3.5	3.7	-1.0	0.3	$11.7_{11.8}$	$11.6_{11.0}$	$10.7_{10.3}$	$10.7_{11.2}$	94	92	95	96	72.1	62.4	49.3	50.6
	AUG_{CMB}	4.5	7.4	-0.9	0.5	$11.6_{11.7}$	$12.2_{11.2}$	$10.4_{10.5}$	$10.6_{11.2}$	95	87	95	95	67.7	22.1	52.6	50.6
	$\text{AUG}_X^{\text{cox}}$	4.8	-2.8	-0.5	0.5	$12.9_{13.6}$	$11.6_{11.0}$	$11.3_{11.5}$	$11.9_{12.5}$	95	93	96	97	36.1	70.9	35.8	20.9
	$\text{AUG}_{X,S}^{\text{cox}}$	2.1	-0.5	-0.7	0.4	$11.9_{11.9}$	$11.4_{10.8}$	$11.0_{10.3}$	$10.9_{11.2}$	95	94	95	95	76.7	86.3	42.8	45.0
	$\text{AUG}_{\text{CMB}}^{\text{cox}}$	1.1	-4.2	-0.4	0.6	$12.4_{11.8}$	$11.5_{10.2}$	$10.8_{10.3}$	$10.8_{11.1}$	95	90	94	95	67.3	62.7	47.6	45.2

Table A.2: Prediction accuracy for models estimated via various procedures when the true underlying $\mathbf{X}$, $T^\dagger$ and S are generated from ($\mathscr{M}_1$) or ($\mathscr{M}_2$) and C generated from ($\mathscr{C}_{\text{KM}}$), ($\mathscr{C}_{\text{cox}}$) or ($\mathscr{C}_{\text{ncox}}$).

Model for $T^\dagger$	Censoring	Measure	IPW_{KM}	IPW_{cox}	AUG_{CMB}	$\text{AUG}_{\text{CMB}}^{\text{cox}}$	Cox
($\mathscr{M}_1$)	($\mathscr{C}_{\text{KM}}$)	IPEV	0.004	0.008	0.015	0.015	0.017
		AUC	0.625	0.627	0.634	0.634	0.635
	($\mathscr{C}_{\text{cox}}$)	IPEV	-0.004	0.011	0.017	0.016	0.02
		AUC	0.576	0.625	0.630	0.633	0.630
	($\mathscr{C}_{\text{ncox}}$)	IPEV	-0.012	0.018	0.021	0.02	0.022
		AUC	0.541	0.633	0.630	0.636	0.630
($\mathscr{M}_1$)	($\mathscr{C}_{\text{KM}}$)	IPEV	0.088	0.085	0.083	0.082	0.042
		AUC	0.751	0.750	0.748	0.748	0.683
	($\mathscr{C}_{\text{cox}}$)	IPEV	0.087	0.084	0.082	0.08	0.035
		AUC	0.756	0.751	0.751	0.750	0.662
	($\mathscr{C}_{\text{ncox}}$)	IPEV	0.088	0.086	0.083	0.083	0.040
		AUC	0.751	0.754	0.746	0.750	0.675

Appendix B

Appendix B for Chapter 2

Notation

Throughout, we define $\mathbb{J} = E[\mathbf{X}_i^{\otimes 2}\dot{g}(\bar{\boldsymbol{\beta}}_\tau^\mathsf{T}\mathbf{X}_i)]$, $\mathbf{F}_{1i} = \mathbb{J}^{-1}\mathbf{X}_i(Y_{\tau i} - g(\bar{\boldsymbol{\beta}}_\tau^\mathsf{T}\mathbf{X}_i))$, $\mathbf{F}_{2i} = \mathbb{J}^{-1}\mathbf{X}_i(Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^\mathsf{T}\boldsymbol{\Phi}_i(\mathbf{X}_i))$, and $\mathbf{F}_{3i} = \mathbb{J}^{-1}\mathbf{X}_i(Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^\mathsf{T}\boldsymbol{\Phi}_i(\mathbf{Z}_i))$. For any matrix $\mathbf{A}$, $\mathbf{A}_{j,}$ represents the j^{th} row vector, and for any a vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\mathsf{T}$.

B.1 Consistency of $\widehat{\boldsymbol{\beta}}$

Under a mild condition that there does not exist a $\boldsymbol{\beta}$ such that $P(\boldsymbol{\beta}^T\mathbf{X}_1 > \boldsymbol{\beta}^T\mathbf{X}_2 | T_1^\dagger \leq t \leq T_2^\dagger) = 1$, using a similar argument given by Tian (2007, app.A), we can show that $\mathbf{U}_0(\boldsymbol{\beta}) = 0$ has a unique solution. To show $\hat{\boldsymbol{\beta}}$ converges to $\bar{\boldsymbol{\beta}}$ in probability, we have to show $\sup_{\boldsymbol{\beta}} |\hat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_0(\boldsymbol{\beta})| = o_p(1)$.

Let $\bar{\boldsymbol{\theta}}_{t_s}$ be the solution to the estimating equation

$$\mathbf{Q}_0(\boldsymbol{\theta}) = E[\Phi(\mathbf{X}_i)(Y_{t_s i} - g(\boldsymbol{\theta}^T\Phi(\mathbf{X}_i)))] = 0$$

Using similar arguments to prove $\tilde{\boldsymbol{\beta}}$ converges to $\bar{\boldsymbol{\beta}}$ in probability and coupled with the fact that $\lambda_{t_s} \to 0$, one can prove that $\widehat{\boldsymbol{\theta}}_{t_s}$ converges to $\bar{\boldsymbol{\theta}}_{t_s}$ in probability by showing weak convergence of $\widehat{\mathbf{Q}}_n(\cdot)$ to $\mathbf{Q}_0(\cdot)$ and assuming some mild regularity conditions.

Let $\bar{\boldsymbol{\gamma}}_\tau$ be the solution to the estimating equation

$$\mathbf{D}_0(\boldsymbol{\gamma}) = E[\Phi(\mathbf{Z}_i)(Y_{\tau i} - g(\boldsymbol{\gamma}^T\Phi(\mathbf{Z}_i)))|T_i^\dagger > t_s] = 0$$

Let $\mathbf{D}_n^*(\boldsymbol{\gamma}) = \frac{1}{\hat{S}(t_s)}\frac{1}{n}\sum_{i=1}^n\{I(T_i > t_s)w_{\tau i}\Phi(\mathbf{Z}_i)(Y_{\tau i} - g(\boldsymbol{\gamma}^\intercal\Phi(\mathbf{Z}_i)))\}$. By the law of large number and consistency of $\hat{S}(t_s)$,

$$\begin{aligned}
\mathbf{D}_n^*(\boldsymbol{\gamma}) &\to \frac{1}{S(t_s)}E\{I(T_i > t_s)w_{\tau i}\Phi(\mathbf{Z}_i)(Y_{\tau i} - g(\boldsymbol{\gamma}^\intercal\Phi(\mathbf{Z}_i)))\} \\
&= \frac{1}{S(t_s)}E[E\{I(T_i > t_s)w_{\tau i}\Phi(\mathbf{Z}_i)(Y_{\tau i} - g(\boldsymbol{\gamma}^\intercal\Phi(\mathbf{Z}_i)))|T_i^\dagger, \mathbf{Z}_i\}] \\
&= \frac{1}{S(t_s)}E[\Phi(\mathbf{Z}_i)(Y_{\tau i} - g(\boldsymbol{\gamma}^\intercal\Phi(\mathbf{Z}_i)))E\{I(T_i > t_s)w_{\tau i}|T_i^\dagger, \mathbf{Z}_i\}] \\
&= \frac{1}{S(t_s)}E[\Phi(\mathbf{Z}_i)(Y_{\tau i} - g(\boldsymbol{\gamma}^\intercal\Phi(\mathbf{Z}_i)))E\{\frac{I(t_s < T_i^\dagger < \tau)I(T_i^\dagger \le C_i)}{G(T_i^\dagger)} \\
&\qquad + \frac{I(T_i^\dagger > \tau)I(C_i > \tau)}{G(\tau)}|T_i^\dagger\}] \\
&= \frac{1}{S(t_s)}E[\Phi(\mathbf{Z}_i)(Y_i - g(\boldsymbol{\gamma}^T\Phi(\mathbf{Z}_i)))\{I(t_s \le T_i^\dagger \le \tau) + I(T_i^\dagger > \tau)\}] \\
&= \frac{1}{S(t_s)}E[I(T_i^\dagger > t_s)\Phi(\mathbf{Z}_i)(Y_i - g(\boldsymbol{\gamma}^T\Phi(\mathbf{Z}_i)))] \\
&= \mathbf{D}_0(\boldsymbol{\gamma})
\end{aligned}$$

The above proof can also be simplified with the fact that $I(T_i > t_s)w_{\tau i} = I(T_i^\dagger > t_s)w_{\tau i}$. By the uniform law of large numbers (Pollard, 1990), that $sup_{\boldsymbol{\gamma}}|\mathbf{D}_n^*(\boldsymbol{\gamma}) - \mathbf{D}_0(\boldsymbol{\gamma})| \to 0$ in probability. The uniform consistency of $\hat{G}(s)$ implies that $sup_{\boldsymbol{\gamma}}|\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) - \mathbf{D}_n^*(\boldsymbol{\gamma})| \to 0$ in probability. Thus $sup_{\boldsymbol{\gamma}}|\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) - \mathbf{D}_0(\boldsymbol{\gamma})| \to 0$ and $\widehat{\boldsymbol{\gamma}}_\tau$ converges to $\bar{\boldsymbol{\gamma}}_\tau$ in probability.

Let $\mathbf{U}_n^*(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{X}_i\{g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i)) + \frac{I(T_i>t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)) - g(\boldsymbol{\beta}^T\mathbf{X}_i)\}$ and $\mathbf{U}_0^*(\boldsymbol{\beta}) = E\{\mathbf{X}_i[g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i)) + \frac{I(T_i>t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)) - g(\boldsymbol{\beta}^T\mathbf{X}_i)]\}$. By law of larger number, $\mathbf{U}_n^*(\boldsymbol{\beta}) \to \mathbf{U}_0^*(\boldsymbol{\beta})$. Moreover, since $|\frac{\partial\mathbf{U}_n^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}| = |n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T\dot{g}(\boldsymbol{\beta}^T\mathbf{X}_i)|$ is bounded, $\sup_{\boldsymbol{\beta}}|U_n^*(\boldsymbol{\beta}) - \mathbf{U}_0^*(\boldsymbol{\beta})| = o_p(1)$.

Note that $\mathbf{U}_0^*(\boldsymbol{\beta}) = U_0(\boldsymbol{\beta})$ because:

$$\begin{aligned}
\mathbf{U}_0(\boldsymbol{\beta}) - \mathbf{U}_0^*(\boldsymbol{\beta}) &= E\{\mathbf{X}[I(T_i^\dagger \le t_s) + I(T_i^* > t_s)I(T_i^\dagger \le \tau) - g(\boldsymbol{\beta}^T\mathbf{X}_i)]\} \\
&\quad - E\{\mathbf{X}_i[g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)) - g(\boldsymbol{\beta}^T\mathbf{X}_i)]\} \\
&= E\{\mathbf{X}_i[Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i))]\} + E\{\mathbf{X}_i I(T_i^\dagger \le \tau)|T_i^\dagger > t_s\}P(T_i^\dagger > t_s) \\
&\quad - E\{\mathbf{X}_i g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i))|T_i > t_s\}\frac{P(T_i > t_s)}{G(t_s)} \\
&= E\{\mathbf{X}_i[Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i))]\} \\
&\quad + E\{\mathbf{X}_i[I(T_i^\dagger \le \tau) - g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i))]|T_i > t_s\}P(T_i^\dagger > t_s)
\end{aligned}$$

The last equality holds because $E\{\mathbf{X}_i I(T_i^* \le \tau)|T_i^\dagger > t_s\} = E\{\mathbf{X}_i I(T_i^\dagger \le \tau)|T_i > t_s\}$ due to independence between C_i and $(\mathbf{X}_i, T_i^\dagger)$. Both terms in the above quantity are 0 because $Q_0(\bar{\boldsymbol{\theta}}_\tau) = D_0(\bar{\boldsymbol{\gamma}}) = 0$ and $\mathbf{X}$ is part of $\Phi(\mathbf{X})$ and $\Phi(\mathbf{Z})$.

We next show that $\sup_{\boldsymbol{\beta}} |\hat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_n^*(\boldsymbol{\beta})| = o_p(1)$:

$$\begin{aligned}
|\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_n^*(\boldsymbol{\beta})| &= |n^{-1}\sum_{i=1}^n \mathbf{X}_i[g(\widehat{\boldsymbol{\theta}}_{t_s}^\intercal\Phi(\mathbf{X}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal\Phi(\mathbf{X}_i)) \\
&\quad + \frac{I(T_i > t_s)}{\hat{G}(t_s)}g(\widehat{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)) - \frac{I(T_i > t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\Phi(\mathbf{Z}_i))]| \\
&\le |n^{-1}\sum_{i=1}^n \mathbf{X}_i[g(\widehat{\boldsymbol{\theta}}_{t_s}^\intercal\Phi(\mathbf{X}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i))]| \\
&\quad + |n^{-1}\sum_{i=1}^n \mathbf{X}_i\frac{I(T_i > t_s)}{\hat{G}(t_s)}g(\widehat{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)) - \frac{I(T_i > t_s)}{G(t_s)}g(\widehat{\boldsymbol{\gamma}}_\tau^\intercal\Phi(\mathbf{Z}_i))]| \\
&\le n^{-1}|(\bar{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s})\sum_{i=1}^n \mathbf{X}_i\Phi(\mathbf{X}_i)^T\dot{g}(\boldsymbol{\theta}_i^*)| \\
&\quad + n^{-1}|(\widehat{\boldsymbol{\gamma}}_\tau - \widehat{\boldsymbol{\gamma}}_\tau)(\frac{1}{\hat{G}(t_s)} - \frac{1}{G(t_s)})\sum_{i=1}^n \mathbf{X}_i\Phi(\mathbf{Z}_i)^T\dot{g}(\boldsymbol{\gamma}_i^*)| \\
&\le |\widehat{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s}|n^{-1}\sum_{i=1}^n |\mathbf{X}_i\Phi(\mathbf{X}_i)^T\dot{g}(\boldsymbol{\theta}_i^*)| \\
&\quad + |\bar{\boldsymbol{\gamma}}_\tau - \bar{\boldsymbol{\gamma}}_\tau||\frac{1}{\hat{G}(t_s)} - \frac{1}{G(t_s)}|n^{-1}\sum_{i=1}^n |\mathbf{X}_i\Phi(\mathbf{X}_i)^T\dot{g}(\boldsymbol{\gamma}_i^*)|
\end{aligned}$$

$$= o_p(1) * O_p(1) + o_p(1) * o_p(1) * O_p(1)$$
$$= o_p(1)$$

It follows that $\sup_{\boldsymbol{\beta}} |\hat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_0^*(\boldsymbol{\beta})| \leq \sup_{\boldsymbol{\beta}} |\hat{\mathbf{U}}_n(\boldsymbol{\beta}) - U_n^*(\boldsymbol{\beta})| + \sup_{\boldsymbol{\beta}} |\mathbf{U}_n^*(\boldsymbol{\beta}) - U_0^*(\boldsymbol{\beta})| = o_p(1)$. Since $\mathbf{U}_0^*(\boldsymbol{\beta}) = U_0(\boldsymbol{\beta})$, we have $\sup_{\beta} |\hat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbf{U}_0(\boldsymbol{\beta})| = o_p(1)$.

B.2 Asymptotic Normality of $\widehat{\boldsymbol{\beta}}$

For the simpler case where $G(\cdot)$ and $G_{t_s}(\cdot)$ (thus $w_{t_s i}$ and $w_{\tau i}$) are known:

The estimating function for $\widehat{\boldsymbol{\theta}}_{t_s}$ is

$$\widehat{\mathbf{Q}}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} w_{t_s i} \boldsymbol{\Phi}(\mathbf{X}_i)\{Y_{t_s i} - g(\boldsymbol{\theta}^T \boldsymbol{\Phi}(\mathbf{X}_i))\} + \lambda_{t_s} \boldsymbol{\theta}$$

if $\lambda_{t_s} = o_p(n^{-\frac{1}{2}})$ and using Taylor expansion :

$$\sqrt{n}\widehat{\mathbf{Q}}_n(\widehat{\boldsymbol{\theta}}_{t_s}) \approx \sqrt{n}\widehat{\mathbf{Q}}_n(\bar{\boldsymbol{\theta}}_{t_s}) + \sqrt{n}\frac{\partial \widehat{\mathbf{Q}}_n(\bar{\boldsymbol{\theta}}_{t_s})}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s})$$
$$0 \approx \sqrt{n}\widehat{\mathbf{Q}}_n(\bar{\boldsymbol{\theta}}_{t_s}) - [\frac{1}{n}\sum_{i=1}^{n} w_{t_s i} \boldsymbol{\Phi}(\mathbf{X}_i)^{\otimes 2} \dot{g}(\bar{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i))]\sqrt{n}(\widehat{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s})$$

Let $\mathbb{A} = \{E[\boldsymbol{\Phi}(\mathbf{X})^{\otimes 2} \dot{g}(\bar{\boldsymbol{\theta}}_{t_s}^{\intercal} \boldsymbol{\Phi}(\mathbf{X}))]\}$. By law of large number,
$\frac{1}{n}\sum_{i=1}^{n} w_{t_s i} \boldsymbol{\Phi}(\mathbf{X}_i)^{\otimes 2} \dot{g}(\bar{\boldsymbol{\theta}}_{t_s}^{\intercal} \boldsymbol{\Phi}(\mathbf{X}_i)) \to \mathbb{A}$, and we can write

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s}) \approx \mathbb{A}^{-1}\sqrt{n}\widehat{\mathbf{Q}}_n(\bar{\boldsymbol{\theta}}_{t_s})$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{t_s i} \mathbb{A}^{-1} \boldsymbol{\Phi}(\mathbf{X}_i)\{Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^{\intercal} \boldsymbol{\Phi}(\mathbf{X}_i)\}$$

The estimating function for $\widehat{\boldsymbol{\gamma}}_\tau$ is

$$\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t_s)\frac{w_{\tau i}}{\hat{S}_{t_s}} \Phi(\mathbf{Z}_i)\{Y_{\tau i} - g(\boldsymbol{\gamma}^{\intercal}\Phi(\mathbf{Z}_i))\} + \lambda_{n\tau}\boldsymbol{\gamma}$$

if $\lambda_{n\tau} = o_p(n^{-\frac{1}{2}})$ and using Taylor expansion :

$$\sqrt{n}\widehat{\mathbf{D}}_n(\widehat{\boldsymbol{\gamma}}_\tau) \approx \sqrt{n}\widehat{\mathbf{D}}_n(\bar{\boldsymbol{\gamma}}_\tau) + \sqrt{n}\frac{\partial\widehat{\mathbf{D}}_n(\bar{\boldsymbol{\gamma}}_\tau)}{\partial\boldsymbol{\gamma}}(\widehat{\boldsymbol{\gamma}}_\tau - \bar{\boldsymbol{\gamma}}_\tau)$$

$$0 \approx \sqrt{n}\widehat{\mathbf{D}}_n(\bar{\boldsymbol{\gamma}}_\tau) - [\frac{1}{n\hat{S}(t_s)}\sum_{i=1}^{n} I(T_i > t_s)w_{\tau i}\boldsymbol{\Phi}(\mathbf{Z}_i)^{\otimes 2}\dot{g}(\bar{\boldsymbol{\gamma}}_\tau^T\boldsymbol{\Phi}(\mathbf{Z}_i))]\sqrt{n}(\widehat{\boldsymbol{\gamma}}_\tau - \bar{\boldsymbol{\gamma}}_\tau)$$

Let $\mathbb{B} = E[I(T_i^\dagger > t_s)\boldsymbol{\Phi}(\mathbf{Z})^{\otimes 2}\dot{g}(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}))]$. By law of large number, $\frac{1}{n}\sum_{i=1}^{n} I(T_i > t_s)w_{\tau i}\boldsymbol{\Phi}(\mathbf{Z}_i)^{\otimes 2}\dot{g}(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)) \to \mathbb{B}$, and we can write

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\gamma}}_\tau - \bar{\boldsymbol{\gamma}}_\tau) &\approx \mathbb{B}^{-1}\sqrt{n}\widehat{\mathbf{D}}_n(\bar{\boldsymbol{\gamma}}_\tau) \\ &\approx \frac{1}{\sqrt{n}}\mathbb{B}^{-1}\sum_{i=1}^{n} I(T_i > t_s)w_{\tau i}\boldsymbol{\Phi}(\mathbf{Z}_i)\{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)\}\end{aligned}$$

The estimating function for $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\{g(\hat{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\hat{\boldsymbol{\gamma}}_\tau^\intercal\Phi(\mathbf{Z}_i)) - g(\boldsymbol{\beta}^\intercal\mathbf{X}_i)\}$$

Using similar arguments above, let $\mathbb{J} = E[\mathbf{X}^{\otimes 2}\dot{g}(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X})]$, we have:

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) &\approx \mathbb{J}^{-1}\sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) \\ &= \mathbb{J}^{-1}\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\{g(\widehat{\boldsymbol{\theta}}_{t_s}^T\boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\widehat{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}_i)\}] \\ &= \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i[\{g(\widehat{\boldsymbol{\theta}}_{t_s}^\intercal\boldsymbol{\Phi}(\mathbf{X}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal\boldsymbol{\Phi}(\mathbf{X}_i))\} \\ &\quad + \frac{I(T_i > t_s)}{G(t_s)}\{g(\widehat{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i))\} \\ &\quad + \{g(\bar{\boldsymbol{\theta}}_{t_s}^T\boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}_i)\}] \\ &\approx \mathbb{J}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{\Phi}(\mathbf{X}_i)^\intercal\dot{g}(\bar{\boldsymbol{\theta}}_{t_s}^\intercal\boldsymbol{\Phi}(\mathbf{X}_i))\sqrt{n}(\widehat{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s}) \\ &\quad + \mathbb{J}^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{I(T_i > t_s)}{G(t_s)}\mathbf{X}_i\boldsymbol{\Phi}(\mathbf{Z}_i)^\intercal\dot{g}(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i))\sqrt{n}(\widehat{\boldsymbol{\gamma}}_\tau - \bar{\boldsymbol{\gamma}}_\tau) \\ &\quad + \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i\{g(\bar{\boldsymbol{\theta}}_{t_s}^T\boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}_i)\}\end{aligned}$$

Let $\mathbb{K} = E[\mathbf{X}\boldsymbol{\Phi}(\mathbf{X})^\intercal \dot{g}(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}))]$ and $\mathbb{L} = E[I(T_i^\dagger > t_s)\mathbf{X}\boldsymbol{\Phi}(\mathbf{Z})^\intercal \dot{g}(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))]$, by the law of large number, $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\boldsymbol{\Phi}(\mathbf{X}_i)^\intercal \dot{g}(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) \to \mathbb{K}$
and $\frac{1}{n}\sum_{i=1}^n \frac{I(T_i>t_s)}{G(t_s)}\mathbf{X}_i\boldsymbol{\Phi}(\mathbf{Z}_i)^\intercal \dot{g}(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) \to \mathbb{L}$.
Substitute $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s}) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^n w_{t_si}\mathbb{A}^{-1}\boldsymbol{\Phi}(\mathbf{X}_i)\{Y_{t_si} - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)\}$ and $\sqrt{n}(\widehat{\boldsymbol{\gamma}}_{t_s} - \bar{\boldsymbol{\gamma}}_{t_s}) \approx \frac{1}{\sqrt{n}}\mathbb{B}^{-1}\sum_{i=1}^n I(T_i > t_s)w_{\tau i}\boldsymbol{\Phi}(\mathbf{Z}_i)\{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)\}$, then the above equation becomes

$$
\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \approx\ & \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n w_{t_si}\mathbb{K}\mathbb{A}^{-1}\boldsymbol{\Phi}(\mathbf{X}_i)\{Y_{t_si} - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)\} \\
& + \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n I(T_i > t_s)w_{\tau i}\mathbb{L}\mathbb{B}^{-1}\boldsymbol{\Phi}(\mathbf{Z}_i)\{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)\} \\
& + \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{g(\bar{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)\} \\
\approx\ & \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n w_{t_si}\mathbf{X}_i[\{Y_{t_si} - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\} + I(T_i > t_s)w_{\tau i}\mathbf{X}_i\{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\} \\
& + \mathbf{X}_i\{g(\bar{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)}g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)\}]
\end{aligned}
$$

Note that

$$
\begin{aligned}
I(T_i^\dagger > t_s)w_{\tau i} + w_{t_si} &= \frac{I(t_s \le T_i^\dagger \le \tau)I(T_i^\dagger \le C_i)}{G(T_i^\dagger)} + \frac{I(T_i \ge \tau)}{G(\tau)} \\
&\quad + \frac{I(T_i^\dagger \le t_s)I(T_i^\dagger \le C_i)}{G(T_i^\dagger)} + \frac{I(T_i \ge t_s)}{G(t_s)} \\
&= w_{\tau i} + \frac{I(T_i \ge t_s)}{G(t_s)}
\end{aligned}
$$

Therefore $w_{t_si} = w_{\tau i} - I(T_i^\dagger > t_s)w_{\tau i} + \frac{I(T_i>t_s)}{G(t_s)}$. Along with the fact $I(T_i^\dagger > t_s)w_{\tau i} = I(T_i > t_s)w_{\tau i}$, the above equation can be written as:

$$
\sqrt{n}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \approx \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{I(T_i^\dagger < t_s)w_{\tau i} + \frac{I(T_i > t_s)}{G(t_s)}\}\mathbf{X}_i[\{Y_{t_si} - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}
$$

$$
\begin{aligned}
&+ I(T_i^\dagger > t_s) w_{\tau i} \mathbf{X}_i \{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\} \\
&+ \mathbf{X}_i \{g(\bar{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)} g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)\}] \\
\approx\ & \mathbb{J}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i [w_{\tau i} \{Y_\tau - I(T_i^\dagger < t_s) g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) - I(T_i^\dagger > t_s) g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\} \\
&- \frac{I(T_i > t_s)}{G(t_s)} g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) + g(\bar{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)} g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)] \\
\approx\ & \mathbb{J}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i [(w_{\tau i} - 1) \{Y_\tau - I(T_i^\dagger < t_s) g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) - I(T_i^\dagger > t_s) g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\} \\
&+ Y_\tau - I(T_i^\dagger < t_s) g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) - I(T_i^\dagger > t_s) g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) \\
&- \frac{I(T_i > t_s)}{G(t_s)} g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) + g(\bar{\boldsymbol{\theta}}_{t_s}^T \boldsymbol{\Phi}(\mathbf{X}_i)) + \frac{I(T_i > t_s)}{G(t_s)} g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)] \\
\approx\ & \mathbb{J}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_i [Y_\tau - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i) \\
&+ (\frac{I(C_i > t_s)}{G(t_s)} - 1) I(T_i^\dagger > t_s) \{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\} \\
&+ (w_{\tau i} - 1) \{Y_\tau - I(T_i^\dagger < t_s) g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i)) - I(T_i^\dagger > t_s) g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\}]
\end{aligned}
$$

Suppose $f(T_i^\dagger, \mathbf{Z}_i)$ is any function of $(T_i^\dagger, \mathbf{Z}_i)$, we have

$$
\begin{aligned}
&E[(w_{\tau i} - 1)(\frac{I(C_i > t_s)}{G(t_s)} - 1) f(T_i^\dagger, \mathbf{Z}_i)] \\
&= E[E\{(w_{\tau i} - 1)(\frac{I(C_i > t_s)}{G(t_s)} - 1) f(T_i^\dagger, \mathbf{Z}_i) | T_i^\dagger, \mathbf{Z}_i\}] \\
&= E[f(T_i^\dagger, \mathbf{Z}_i) E\{(w_{\tau i} - 1)(\frac{I(C_i > t_s)}{G(t_s)} - 1) | T_i^\dagger, \mathbf{Z}_i\}] \\
&= E[f(T_i^\dagger, \mathbf{Z}_i) E\{(w_{\tau i} - 1)\frac{I(C_i > t_s)}{G(t_s)} | T_i^\dagger, \mathbf{Z}_i\}] \\
&= E[f(T_i^\dagger, \mathbf{Z}_i) E\{(\frac{I(T_i^\dagger \le C_i) I(T_i^\dagger < \tau)}{G(T_i^\dagger)} \\
&\quad + \frac{I(T_i^\dagger > \tau) I(C_i > \tau)}{G(\tau)} - 1)\frac{I(C_i > t_s)}{G(t_s)} | T_i^\dagger, \mathbf{Z}_i\}]
\end{aligned}
$$

$$
\begin{aligned}
&= E[f(T_i^\dagger, \mathbf{Z}_i)\frac{I(T_i^\dagger < \tau)}{G(T_i^\dagger)G(t_s)}E\{I(T_i^\dagger \le C_i)I(C_i > t_s)|T_i^\dagger, \mathbf{Z}_i\} \\
&\quad + E[f(T_i^\dagger, \mathbf{Z}_i)\frac{I(T_i^\dagger > \tau)}{G(\tau)G(t_s)}E\{I(C_i > \tau)|T_i^\dagger, \mathbf{Z}_i\} - E[f(T_i^\dagger, \mathbf{Z}_i)] \\
&= E[f(T_i^\dagger, \mathbf{Z}_i)\{\frac{I(T_i^\dagger < t_s)}{G(T_i^\dagger)} + \frac{I(t_s \le T_i^\dagger \le \tau)}{G(t_s)} + \frac{I(T_i^\dagger > \tau)}{G(t_s)} - 1\}] \\
&= E[f(T_i^\dagger, \mathbf{Z}_i)\{\frac{I(T_i^\dagger < t_s)}{G(T_i^\dagger)} + \frac{I(T_i^\dagger > t_s)}{G(t_s)} - 1\}]
\end{aligned}
$$

It is also easy to see that $E[(w_{\tau i}-1)(\frac{I(C_i>t_s)}{G(t_s)}-1)I(T_i^\dagger > t_s)f(T_i^\dagger, \mathbf{Z}_i)] = E[(\frac{1}{G(t_s)}-1)I(T_i^\dagger > t_s)f(T_i^\dagger, \mathbf{Z}_i)]$. Therefore, by CLT, we have $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \to N(0, \hat{\boldsymbol{\Sigma}})$ where

$$
\begin{aligned}
&\hat{\boldsymbol{\Sigma}} \\
&= Var[\mathbf{F}_1] + Var[\mathbb{J}^{-1}\mathbf{X}_i(\frac{I(C_i > t_s)}{G(t_s)} - 1)I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}] \\
&\quad + Var[\mathbb{J}^{-1}\mathbf{X}_i(w_{\tau i}-1)\{Y_\tau - I(T_i^\dagger > t_s)g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - I(T_i^\dagger < t_s)g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}] \\
&\quad + E[\{\frac{1}{G(t_s)} - 1\}I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&= Var[\mathbf{F}_1] + E[\{\frac{I(C_i > t_s)}{G(t_s)} - 1)\}^2]E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}^2(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&\quad + E[\{w_{\tau i}-1\}^2\{Y_\tau - I(T_i^\dagger > t_s)g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - I(T_i^\dagger < t_s)g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}^2\{\mathbb{J}^{-1}\mathbf{X}_i\}^{\otimes 2}] \\
&\quad + E[\{\frac{1}{G(t_s)} - 1\}I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&= Var[\mathbf{F}_1] + (\frac{1}{G(t_s)} - 1)E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}^2(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&\quad + E[\{\int_0^\tau I(T_i^\dagger > s)d\frac{1}{G(s)}\}\{Y_\tau - I(T_i^\dagger > t_s)g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) \\
&\quad - I(T_i^\dagger < t_s)g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}^2\{\mathbb{J}^{-1}\mathbf{X}_i\}^{\otimes 2}] \\
&\quad + \{\frac{1}{G(t_s)} - 1\}E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&= Var[\mathbf{F}_1] + E[\{\int_0^\tau I(T_i^\dagger > s)d\frac{1}{G(s)}\}\{I(T_i^\dagger < t_s)\mathbf{F}_{2i}^{\otimes 2} + I(T_i^\dagger > t_s)\mathbf{F}_{3i}^{\otimes 2}\}] \\
&\quad \{\frac{1}{G(t_s)} - 1\}E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}]
\end{aligned}
$$

Zheng and Cai (2017) has shown that the asymptotic variance of $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\Sigma}}$, is $Var[\mathbf{F}_{1i}] + Var[(w_{\tau i}-1)\mathbf{F}_{1i}]$ and $Var[(w_{\tau i}-1)\mathbf{F}_{1i}] = E[\{\int_0^\tau I(T_i^\dagger > s)d\frac{1}{G(s)}\}\mathbf{F}_{1i}^{\otimes 2}]$. Therefore,

the variance reduction is

$$
\begin{aligned}
&\tilde{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}} \\
&= E[\{\int_0^\tau I(T_i^\dagger > s) d\frac{1}{G(s)}\}\mathbf{F}_{1i}^{\otimes 2}] \\
&\quad - E[\{\int_0^\tau I(T_i^\dagger > s) d\frac{1}{G(s)}\}\{I(T_i^\dagger < t_s)\mathbf{F}_{2i}^{\otimes 2} + I(T_i^\dagger > t_s)\mathbf{F}_{3i}^{\otimes 2}\}] \\
&\quad - \{\frac{1}{G(t_s)} - 1\}E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \mathbf{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&= E[\{\int_0^{t_s} I(T_i^\dagger > s) d\frac{1}{G(s)}\}\mathbf{F}_{1i}^{\otimes 2}] + E[\{\int_{t_s}^\tau I(T_i^\dagger > s) d\frac{1}{G(s)}\}\mathbf{F}_{1i}^{\otimes 2}] \\
&\quad - E[\{\int_0^{t_s} I(T_i^\dagger > s) d\frac{1}{G(s)}\}\{I(T_i^\dagger < t_s)\mathbf{F}_{2i}^{\otimes 2} + I(T_i^\dagger > t_s)\mathbf{F}_{3i}^{\otimes 2}\}] \\
&\quad - E[\{\int_{t_s}^\tau I(T_i^\dagger > s) d\frac{1}{G(s)}\}\mathbf{F}_{3i}^{\otimes 2}] \\
&\quad - \{\frac{1}{G(t_s)} - 1\}E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \mathbf{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&= E[\int_{t_s}^\tau \{I(T_i^\dagger > s)(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{3i}^{\otimes 2})\} d\frac{1}{G(s)}] + E[\int_0^{t_s} \{I(T_i^\dagger > s)\mathbf{F}_{1i}^{\otimes 2}\} d\frac{1}{G(s)}] \\
&\quad - E[\int_0^{t_s} \{I(s < T_i^\dagger < t_s)\mathbf{F}_{2i}^{\otimes 2} + I(T_i^\dagger > t_s)\mathbf{F}_{3i}^{\otimes 2}\} d\frac{1}{G(s)}] \\
&\quad - \{\frac{1}{G(t_s)} - 1\}E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \mathbf{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}] \\
&= E[\int_{t_s}^\tau \{I(T_i^\dagger > s)(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{3i}^{\otimes 2})\} d\frac{1}{G(s)}] + E[\int_0^{t_s} \{I(T_i^\dagger > s)(\mathbf{F}_{1i}^{\otimes 2} - \mathbf{F}_{2i}^{\otimes 2})\} d\frac{1}{G(s)}] \\
&\quad - \{\frac{1}{G(t_s)} - 1\}E[I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal \mathbf{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}\{Y_\tau - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \mathbf{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}]
\end{aligned}
\tag{B.1}
$$

Now for the case where $G(T_i \wedge \tau)$ is unknown and estimated by a consistent estimator $\hat{G}(T_i \wedge \tau)$:

The estimating function for $bb\tilde{e}ta$ is $\tilde{\mathbf{U}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \hat{w}_i \mathbf{X}_i\{Y_{\tau i} - g(\boldsymbol{\beta}^T\mathbf{X}_i)\}$. Uno

et.al (2007) has shown that:

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}_\tau) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{J}^{-1}\{w_i\mathbf{X}_i(Y_{\tau i}-g(\bar{\boldsymbol{\beta}}_\tau^T\mathbf{X}_i))+\int_0^\tau \psi_i(s)dE[\mathbf{X}_i(Y_\tau-g(\bar{\boldsymbol{\beta}}_\tau^T\mathbf{X}_i))I(T_i^\dagger\le s)]\}$$

where $\psi_i(s) = \int_0^s \frac{dM_{ci}(u)}{\pi(u)}$ with $\pi(u) = Pr(T_i > u)$, $M_{ci}(u) = I(T_i \le u, \delta_i = 0) - \int_0^u I(T_i > v)d\Lambda_c(v)$ and $\Lambda_c(\cdot)$ is the cumulative hazard function for the censoring variable C. Zheng and Cai (2017) has shown that $\sqrt{n}(\tilde{bbeta} - \bar{\boldsymbol{\beta}}_\tau) \to N(0, \tilde{\Sigma})$ with

$$\tilde{\Sigma} = Var\{\mathbf{F}_{1i}\} + Var\{(w_{\tau i}-1)\mathbf{F}_{1i}\} + \int_0^\tau \{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2} - \boldsymbol{\mu}_{F_1}(s)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}$$

Similarly, it can be shown that

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}}_{t_s} - \bar{\boldsymbol{\theta}}_{t_s}) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{J}^{-1}\{&w_{t_s i}\boldsymbol{\Phi}(\mathbf{X}_i)(Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^T\boldsymbol{\Phi}(\mathbf{X}_i)))\\ &+\int_0^{t_s}\psi_i(s)dE[\boldsymbol{\Phi}(\mathbf{X}_i)(Y_{t_s} - g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i)))I(T_i^\dagger\le s)]\}\end{aligned}$$

and

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\gamma}}_\tau - \bar{\boldsymbol{\gamma}}_\tau) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{J}^{-1}\{&w_\tau\boldsymbol{\Phi}(\mathbf{Z}_i)(Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^T\boldsymbol{\Phi}(\mathbf{Z}_i)))\\ &+\int_{t_s}^{\tau}\psi_i(s)dE[\boldsymbol{\Phi}(\mathbf{Z}_i)(Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)))I(T_i^\dagger\le s)]\}\end{aligned}$$

Consequently,

$$\begin{aligned}&\sqrt{n}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})\\ &\approx \mathbb{J}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i[Y_\tau - g(\bar{\boldsymbol{\beta}}^\intercal\mathbf{X}_i) + (\frac{I(C_i > t_s)}{G(t_s)} - 1)I(T_i^\dagger > t_s)\{g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i)) - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal\boldsymbol{\Phi}(\mathbf{X}_i))\}\\ &\quad + (w_{\tau i}-1)\{Y_\tau - I(T_i^\dagger < t_s)g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal\boldsymbol{\Phi}(\mathbf{X}_i)) - I(T_i^\dagger > t_s)g(\bar{\boldsymbol{\gamma}}_\tau^\intercal\boldsymbol{\Phi}(\mathbf{Z}_i))\}\\ &\quad + \int_0^{t_s}\psi_i(s)dE\{\mathbf{X}_i(Y_{t_s} - g(\bar{\boldsymbol{\theta}}_{t_s}^T\Phi(\mathbf{X}_i)))I(T_i^\dagger\le s)\}\\ &\quad + \int_{t_s}^{\tau}\psi_i(s)dE\{\mathbf{Z}_i(Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_\tau^T\Phi(\mathbf{Z}_i)))I(T_i^\dagger\le s)\}]\end{aligned}$$

By CLT, we have $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \to N(0, \hat{\boldsymbol{\Sigma}})$ where

$$\hat{\boldsymbol{\Sigma}} = Var[\mathbf{F}_1] + E[\{\int_0^\tau I(T_i^\dagger > s)d\frac{1}{G(s)}\}\{I(T_i^\dagger < t_s)\mathbf{F}_{2i}^{\otimes 2} + I(T_i^\dagger > t_s)\mathbf{F}_{3i}^{\otimes 2}\}]$$

$$
\begin{aligned}
&+\int_0^{t_s}\{\boldsymbol{\mu}_{F_2}(t_s)^{\otimes 2}-\boldsymbol{\mu}_{F_2}(s)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}+\int_{t_s}^{\tau}\{\boldsymbol{\mu}_{F_3}(\tau)^{\otimes 2}-\boldsymbol{\mu}_{F_3}(s)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}\\
&\{\frac{1}{G(t_s)}-1\}E[I(T_i^{\dagger}>t_s)\{g(\bar{\boldsymbol{\gamma}}_{\tau}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z}_i))-g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_{\tau}-g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}]\\
=&E[\int_0^{t_s}\{I(T_i^{\dagger}>s)(\mathbf{F}_{1i}^{\otimes 2}-\mathbf{F}_{2i}^{\otimes 2})\}d\frac{1}{G(s)}]+E[\int_{t_s}^{\tau}\{I(T_i^{\dagger}>s)(\mathbf{F}_{1i}^{\otimes 2}-\mathbf{F}_{3i}^{\otimes 2})\}d\frac{1}{G(s)}]\\
&+\int_0^{t_s}\{\boldsymbol{\mu}_{F_2}(t_s)^{\otimes 2}-\boldsymbol{\mu}_{F_2}(s)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}+\int_{t_s}^{\tau}\{\boldsymbol{\mu}_{F_3}(\tau)^{\otimes 2}-\boldsymbol{\mu}_{F_3}(s)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}\\
&\{\frac{1}{G(t_s)}-1\}E[I(T_i^{\dagger}>t_s)\{g(\bar{\boldsymbol{\gamma}}_{\tau}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z}_i))-g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_{\tau}-g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}]\\
=&\int_0^{t_s}[Var\{\mathbf{F}_{1i}|T_i^{\dagger}>s\}-Var\{\mathbf{F}_{2i}|T_i^{\dagger}>s\}]\frac{S(s)^2d\Lambda_c(s)}{\pi(s)}\\
&+\int_0^{t_s}\{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2}-\boldsymbol{\mu}_{F_2}(\tau)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}\\
&\int_{t_s}^{\tau}[Var\{\mathbf{F}_{1i}|T_i^{\dagger}>s\}-Var\{\mathbf{F}_{3i}|T_i^{\dagger}>s\}]\frac{S(s)^2d\Lambda_c(s)}{\pi(s)}\\
&+\int_{t_s}^{\tau}\{\boldsymbol{\mu}_{F_1}(\tau)^{\otimes 2}-\boldsymbol{\mu}_{F_3}(\tau)^{\otimes 2}\}\frac{d\Lambda_c(s)}{\pi(s)}\\
&\{\frac{1}{G(t_s)}-1\}E[I(T_i^{\dagger}>t_s)\{g(\bar{\boldsymbol{\gamma}}_{\tau}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{Z}_i))-g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{X}_i))\}\{Y_{\tau}-g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}(\mathbf{X}_i))\}(\mathbb{J}^{-1}\mathbf{X}_i)^{\otimes 2}]
\end{aligned}
$$

References

Bai, X., Tsiatis, A. A., Lu, W., and Song, R. (2016). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Anal* **22,** 280–298.

Bai, X., Tsiatis, A. A., and OBrien, S. M. (2013). Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics* **69,** 830–839.

Bang, H. and Tsiatis, A. (2000). Esitimating medical cost with censored data. *Biometrika* **87,** 329–343.

Bonetti, M. and Gelber, R. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatitics* **5(3),** 465–81.

Cai, T. and Cheng, S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* **9,** 216–233.

Cai, T., Tian, L., Uno, H., Solomon, S. D., and Wei, L. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika* **97,** 389–404.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 187–220.

DiRienzo, G. (2009). Flexible regression model selection for survival probabilities: with application to aids. *Biometrics* **65,** 1194–1202.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1.

Gerds, T. A., Cai, T., and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical journal* **50,** 457–479.

Gray, R. (1994). A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* **81,** 527–539.

Hammer, S., Katzenstein, D., Hughes., M., Gundacher, H., Schooley, R., R., H., and et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335,** 1081–1090.

Jiang, R., Lu, W., Song, R., and Davidian, M. (2016). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society Series B* **na,** na–na.

Lin, D., Wei, L., and Ying, Z. (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association* **96,** 620–628.

Lu, X. and Tsiatis, A. (2008). Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrika* **95,** 679–694.

Matsouaka, R., Li, J., and Cai, T. (2014). Evaluating marker guided treatment selection strategies. *Biometrics* **70(3),** 489–499.

Newey, W. K. and McFadden, D. (1994). Estimation in large samples. *The Handbook of Econometrics* **4,**.

Parast, L., Tian, L., and Cai, T. (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of American Statistical Association* **109(505),** 384–394.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional Conference series in probability and statistics*, pages i–86. JSTOR.

Robins, J., A., R., and Zhao, L. (1994). Estimation of regression coefficients when some of the regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Robins, J. and Ritov, Y. (1997). A curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16,** 285–319.

Robins, J., Seud, M., Q., L.-G., and Rotnitzky, A. (2007). Comment: performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statistical Science* **22(4),** 544–559.

Robins, J. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87,** 113–124.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, pages 297–331. Springer.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94(448),** 1096–1120.

Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biostatitics* **60(4),** 874–883.

Tsiatis, A. (2006). *Semiparametric Theory and Missing data.* Springer, New York.

Uno, H., Cai, T., Tian, L., and Wei, L. (2007). Evaluating predictionrules for t-year survivors with censored regresion models. *Journal of the American Statistical Association* **102,** 527–537.

van Houwelingen, H. and Putter, H. (2011). *Dynamic Predicitonin Clinical Survival Analysis*. Chapmen & Hall, Florida.

Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34,** 70–85.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94,** 691–703.

Zhao, Y. Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **102(1),** 151168.

Zheng, Y. and Cai, T. (2017). Augmented estimation for t-year survival with censored regression model. *Biometrics* **1,** 00–00.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.