Statistica Sinica Preprint No: SS-2019-0066

Title	Efficient and Robust Estimation of τ-year Risk Prediction Models Leveraging Time Varying Intermediate Outcomes
Manuscript ID	SS-2019-0066
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0066
Complete List of Authors	Yu Zheng Tian Lu and Tianxi Cai
Corresponding Author	Tianxi Cai
E-mail	tcai@hsph.harvard.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Efficient and Robust Estimation of τ-year Risk Prediction Models Leveraging Time Varying Intermediate Outcomes

Yu Zheng[⋆], Tian Lu[†], Tianxi Cai[⋆]

[⋆] *Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115*

[†] *Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305*

Abstract:

Developing accurate risk prediction model is a key step towards 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 predict the τ-year risk of developing a clinical event using baseline covariates. Such τ-year risk models can be estimated by fitting standard survival models including the Cox proportional hazards model and the more flexible τ-year specific generalized linear model (τ-GLM). However, efficient and robust estimation of the risk model is challenging under heavy censoring and potential model mis-specification. Intermediate outcomes observed prior to loss to follow up can be highly predictive of the outcome and thus may be used to improve the efficiency of the model estimation. However, existing augmentation methods either do not allow intermediate outcomes to be subject to censoring or have limited efficiency gain. 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 Diabetes Prevention Program, a randomized clinical trial on high-risk subjects.

Key words and phrases: Efficiency augmentation, Intermediate outcomes, Model mis-specification, Risk prediction, Robustness, Survival.

1. Introduction

Developing accurate risk prediction models is an important task in translational medicine research. Disease prevention and treatment strategies can be tailored towards individual patients according to risks predicted from such models. For disease prognosis and prevention, it is often of interest to predict the τ-year risk of experiencing a clinical event using baseline clinical and biomarker information. Such τ-year risk models can be estimated by fitting a wide range of survival models including the Cox proportional hazards model (Cox, 1972) and the more flexible τ-year specific generalized linear model (τ-GLM) (Uno et al., 2007). However, efficient and robust estimation of the risk model is challenging under heavy censoring and possible model mis-specification. Under model mis-specification, the partial likelihood estimator for the Cox model converges to a quantity that depends on the censoring distribution (Van Houwelingen, 2007; Cai and Cheng, 2008), leading to reproducibility issues since censoring distribution is almost always study dependent. To derive a robust risk model, Uno et al. (2007) proposed an inverse probability weighted (IPW) estimator for τ-GLM such that the model parameters are always convergent to meaningful quantities that are free of censoring distribution. However, the IPW estimator suffers from low efficiency in heavy censoring settings since it discards information from subjects who are censored before τ .

To improve estimation efficiency under general survival settings, various augmentation procedures have been proposed in the literature to leverage auxiliary baseline covariates or intermediate outcomes. For example, Robins, Rotnitzky, and Zhao (1994) employed alternative estimators for the censoring weights to improve the efficiency of IPW estimators. The doubly robust augmented IPW (AIPW) methods provides protection against misspecification of weights and could potentially improve the estimation efficiency by further employing outcome imputations (Scharfstein, Rotnitzky, and Robins, 1999; Bang, 2005; Tsiatis, 2006). DiRienzo (2009) incorporated AIPW method to estimate the τ-GLM where an estimating function involving outcome imputation is augmented to achieve doubly robustness. However, the AIPW estimators may attain little or even negative efficiency gain when the outcome model is mis-specified. In addition, these existing methods tend to perform poorly when the number of baseline covariates and intermediate outcomes is not small. Zhang and Cai (2017) proposed a two-step imputation based procedure that incorporates auxiliary information including post-baseline outcomes to improve the efficiency. The method requires the auxiliary predictors to be fully observed. However, in cohort studies or clinical trials, post baseline intermediate outcomes are often not observable after subjects either experience the primary outcome or censoring. It is not straightforward to adapt the method to the present setting of τ-GLM estimation with the additional complication of intermediate outcome being missing for those who are no longer at risk.

In this paper, we propose robust imputation based methods to improve the estimation of the τ-GLM model parameters that can effectively incorporate intermediate outcomes that are subject to censoring while allowing both the τ-GLM and the imputation models to be mis-specified. Our method can also easily employ regularization to control for overfitting when the number of augmentation variables is not small. When the post-baseline covariates are measured at multiple time points, we further develop a systematic approach to optimally combine several estimators to maximize efficiency. The rest of the manuscript is organized as follows. Section 2 details the estimation and inference procedure. Section 3 presents simulation results demonstrating the consistency and the efficiency gain of the proposed estimator. Section 4 illustrates the proposed method using data from the Diabetes Prevention Program, a placebo-controlled randomized clinical trial investigating whether the change of lifestyle or taking metformin will prevent type 2 diabetes among high-risk adults. Concluding remarks are given in section 5.

2 Methods

Let $T^\dagger$ be a continuous failure time, and $\mathbf{X} = (X_1 = 1, X_2, ..., X_p)^\intercal$ be a $p \times 1$ vector of bounded baseline predictors. Our goal is to develop an accurate and robust risk prediction model for $Y_\tau = I(T^\dagger \le \tau)$ at some pre-specified time τ based on $\mathbf{X}$. We propose to construct the prediction model for Y_τ by fitting the following τ-GLM *working* model:

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

where $g(\cdot)$ is a known smooth probability distribution function and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters. For simplicity, we focus on the logistic link with $g(x) = e^x/(1+e^x)$ throughout although the procedure can be easily modified to accommodate other link functions. We allow $\boldsymbol{\beta}$ to depend on τ but suppress τ from notational ease.

In addition to the event time and baseline covariates, a q-dimensional intermediate outcomes, denoted by $\mathbf{S}$, are collected over time. Without loss of generality, we assume that $\mathbf{S}$ is measured at K visit times, $0 < t_1 < \cdots < t_K < \tau$, and let $\vec{\mathbf{S}} = (\mathbf{S}_{t_1}^{\mathsf{T}}, ..., \mathbf{S}_{t_K}^{\mathsf{T}})^{\mathsf{T}}$, where $\mathbf{S}_t$ denotes $\mathbf{S}$ measured at time t. Due to censoring, for $T^{\dagger}$, we only observe $T = \min(T^{\dagger}, C)$ and $\delta = I(T^{\dagger} \leq C)$, where C is the censoring time assumed independent of $(T^{\dagger}, \vec{\mathbf{S}}^{\mathsf{T}}, \mathbf{X}^{\mathsf{T}})$ with a common survival function $G(\cdot)$. We allow $\mathbf{S}_t$ to be missing for those who have censored or experienced the event by t but assume that $\mathbf{S}_t$ is observable when $T > t$. The underlying data consists of n independent and identically distributed random vectors, $\mathscr{F} = \{(T_i^{\dagger}, C_i, \mathbf{X}_i^{\mathsf{T}}, \vec{\mathbf{S}}_i^{\mathsf{T}}), i = 1, ..., \mathrm{n}\}$, while the observed data consists of $\mathscr{D} = \{(T_i, \delta_i, \mathbf{X}_i^{\mathsf{T}}, \vec{\mathbf{S}}_{T_i-}^{\mathsf{T}}), i = 1, ..., \mathrm{n}\}$, where $\vec{\mathbf{S}}_{T_i-}$ consists of the subvector of $\vec{\mathbf{S}}_i$ that are measured prior to T_i.

2.1 Estimation procedure

To estimate $\boldsymbol{\beta}$ under τ-GLM given in (2.1), we let $\bar{\boldsymbol{\beta}}$ denote the unique solution to

$$\mathbf{U}_0(\boldsymbol{\beta}) = E[\mathbf{X}\{Y_\tau - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X})\}] = 0.$$

When (2.1) is correctly specified, $\bar{\boldsymbol{\beta}}$ is the true model parameter. Under mild model mis-specification, the resulting risk score $\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}$ was shown to approximately maximize a weighted area under the receiver operating characteristic curve among all functions of $\mathbf{X}$ for classifying Y_τ (Eguchi and Copas, 2002). Thus, $\bar{\boldsymbol{\beta}}$ is a sensible target parameters regardless the adequacy of τ-GLM. We aim to derive a τ-year risk model by constructing a consistent estimator of $\bar{\boldsymbol{\beta}}$.

To account for censoring, Uno et al. (2007) proposed an IPW estimator, $\widetilde{\boldsymbol{\beta}}$, as the solution to

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

where $\widehat{w}_{\tau i} = I(T_i \le \tau)\delta_i + I(T_i > \tau)/\widehat{G}(T_i \wedge \tau)$, and $\widehat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$. For the logistic link $g(\cdot)$, $\widetilde{\boldsymbol{\beta}}$ is also the minimizer of the weighted negative logistic log-likelihood

$$\sum_{i=1}^{n} \widehat{w}_{\tau i}\ell(Y_{\tau i}, \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i), \quad \text{where} \quad \ell(y,x) = -y\log\{g(x)\} - (1-y)\log\{1-g(x)\}.$$

Although $\widetilde{\boldsymbol{\beta}} \to \bar{\boldsymbol{\beta}}$ in probability regardless the adequacy of (2.1), it suffers from low efficiency in settings with heavy censoring since it discards information from subjects censored before τ. We propose to derive a more efficient estimator of $\bar{\boldsymbol{\beta}}$ by leveraging the observed information on $\vec{\mathbf{S}}$.

S measured at a single visit We first consider $\mathbf{S}$ measured at a single time point $t_s < \tau$, $\mathbf{S}_{t_s}$, and write

$$Y_\tau = I(T^\dagger \le t_s) + I(t_s < T^\dagger \le \tau) = Y_{t_s} + I(T^\dagger > t_s)Y_\tau.$$

We propose to estimate $\bar{\boldsymbol{\beta}}$ by separately imputing the missing Y_{t_s} and $I(T^\dagger > t_s)Y_\tau$. To this end, let $\mathbf{Z} = (\mathbf{X}^\intercal, \mathbf{S}_{t_s}^\intercal)^\intercal$ where we suppress t_s from $\mathbf{Z}$ for notational ease. For both $\mathbf{X}$ and $\mathbf{Z}$, we consider their possibly non-linear basis functions, $\boldsymbol{\Phi}(\mathbf{X})$ and $\boldsymbol{\Psi}(\mathbf{Z})$, to account for potential non-linear effects, where we let the first p elements of $\boldsymbol{\Phi}(\mathbf{X})$ and $\boldsymbol{\Psi}(\mathbf{Z})$ to be $\mathbf{X}$.

To impute Y_{t_s}, we fit a working model $P(Y_{t_s} = 1 \mid \mathbf{X}) = g\{\boldsymbol{\theta}_{t_s}^\intercal \boldsymbol{\Phi}(\mathbf{X})\}$ and estimate $\boldsymbol{\theta}_{t_s}$ as $\widehat{\boldsymbol{\theta}}_{t_s}$, the minimizer of the penalized IPW likelihood,

$$\widehat{\mathrm{Q}}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n} \widehat{w}_{t_s i}\ell\left\{Y_{t_s i}, \boldsymbol{\theta}^\intercal \boldsymbol{\Phi}_i\right\} + \lambda_1 \mathcal{Q}(|\boldsymbol{\theta}_{[-1]}|), \tag{2.3}$$

where $\boldsymbol{\Phi}_i = \boldsymbol{\Phi}(\mathbf{X}_i)$, $\mathcal{Q}(\cdot)$ is a penalty function such as the ridge or LASSO (Friedman, Hatie, and Tibshirani, 2001) to allow the dimension of $\boldsymbol{\Phi}(\mathbf{X})$ not small relative to n, $\lambda_1 = o(n^{-\frac{1}{2}})$ is a non-negative penalty parameter that controls the degree of regularization, and for any vector $\boldsymbol{a}$, $\boldsymbol{a}_{[-1]}$ represents the subvector of $\boldsymbol{a}$ with its first element removed. We choose the small penalty parameter to reduce the potential bias in the estimated $\widehat{\boldsymbol{\theta}}_{t_s}$.

For $I(T^\dagger > t_s)Y_\tau$, we impose a working model

$$P(Y_\tau = 1 \mid \mathbf{Z}, T^\dagger > t_s) = g\{\boldsymbol{\gamma}_{\tau|t_s}^\intercal \boldsymbol{\Psi}(\mathbf{Z})\}.$$

and use those with $T > t_s$ to estimate $\boldsymbol{\gamma}_{\tau|t_s}$ since $P(Y_\tau = 1 \mid \mathbf{Z}, T^\dagger > t_s) = P(Y_\tau = 1 \mid \mathbf{Z}, T > t_s)$ under independent censoring. For subjects with $T > t_s$, their intermediate outcome information $\mathbf{S}_{t_s}$ and hence $\mathbf{Z}$ are fully observed. We estimate $\boldsymbol{\gamma}_{\tau|t_s}$ as $\widehat{\boldsymbol{\gamma}}_{\tau|t_s}$, the minimizer of an IPW penalized log-likelihood associated with Y_τ among $T_i > t_s$:

$$\widehat{\mathbf{D}}_n(\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n} I(T_i > t_s)\widehat{w}_{\tau i}\ell\left(Y_{t_s i}, \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\Psi}_i\right) + \lambda_2\mathcal{Q}(|\boldsymbol{\gamma}_{[-1]}|) \qquad (2.4)$$

where $\boldsymbol{\Psi}_i = \boldsymbol{\Psi}(\mathbf{Z}_i)$ and $\lambda_2 = o(n^{-\frac{1}{2}})$ is a non-negative penalty parameter.

Combining estimates from these two working models and noting that the expectation of $\varpi_{t_s i} = I(T_i > t_s)/G(t_s)$ given $\mathbf{Z}_i$ and $T_i^\dagger$ is $I(T_i^\dagger > t_s)$, we impute Y_τ as

$$\widehat{Y}_{\tau i}^{t_s} = g(\widehat{\boldsymbol{\theta}}_{t_s}^\mathsf{T}\boldsymbol{\Phi}_i) + \widehat{\varpi}_{t_s i}\ g(\widehat{\boldsymbol{\gamma}}_{\tau|t_s}^\mathsf{T}\boldsymbol{\Psi}_i). \quad \text{where} \quad \widehat{\varpi}_{t_s i} = \frac{I(T_i > t_s)}{\widehat{G}(t_s)}.$$

With the imputed outcome, we now use all subjects in the dataset to estimate $\bar{\boldsymbol{\beta}}$ as $\widehat{\boldsymbol{\beta}}$, the solution to the estimating equation

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) \equiv n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\left\{\widehat{Y}_{\tau i}^{t_s} - g(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\right\} = 0. \qquad (2.5)$$

We show in Supplemental material Appendix A that $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\bar{\boldsymbol{\beta}}$, regardless the adequacy of the τ-GLM or the imputation models. This demonstrates the robustness of the proposed imputation based procedure in that $\widehat{\boldsymbol{\beta}}$ is valid even if both imputation model and the

τ-GLM are mis-specified. On the contrary, under mis-specification of τ-GLM, separately fitting the GLM to Y_τ and Y_{t_s} will likely yield different estimates of the covariate effects. In Supplemental material Appendix B, we show that $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})$ converges in distribution to a multivariate normal with mean zero and covariance matrix

$$\begin{aligned}\boldsymbol{\Sigma}_{t_s} &= \text{var}(\mathbf{F}_{1i}) + \int_0^{t_s} \text{var}(\mathbf{F}_{2i}+\mathbf{L}_i|T_i^\dagger > s)\frac{S(s)^2 d\Lambda_c(s)}{\pi(s)} \\ &\quad + \int_{t_s}^{\tau} \text{var}(\mathbf{F}_{3i}|T_i^\dagger > s)\frac{S(s)^2 d\Lambda_c(s)}{\pi(s)},\end{aligned}$$

where $\mathbf{F}_{1i} = \mathbb{J}^{-1}\mathbf{X}_i\{Y_{\tau i} - g(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)\}$, $\mathbf{F}_{2i} = \mathbb{J}^{-1}\mathbf{X}_i\{Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^\intercal \boldsymbol{\Phi}_i)\}$, $\mathbf{F}_{3i} = \mathbb{J}^{-1}\mathbf{X}_i\{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_{\tau|t_s}^\intercal \boldsymbol{\Psi}_i)\}$, $\mathbf{L}_i = \mathbb{J}^{-1}\mathbf{X}_i^\intercal g(\bar{\boldsymbol{\gamma}}_{\tau|t_s}^\intercal \boldsymbol{\Psi}_i) I(T_i^\dagger > t_s)$, $\pi(t) = P(T_i \geq t)$, $\mathbb{J} = E\{\mathbf{X}_i^{\otimes 2}\dot{g}(\bar{\boldsymbol{\beta}}^\intercal \mathbf{X}_i)\}$, $\bar{\boldsymbol{\theta}}_{t_s}$ and $\bar{\boldsymbol{\gamma}}_{\tau|t_s}$ are the respective limits of $\widehat{\boldsymbol{\theta}}_{t_s}$ and $\widehat{\boldsymbol{\gamma}}_{\tau|t_s}$, $S(t) = P(T_i^\dagger \geq t)$, and $\Lambda_c(\cdot) = -\log\{G(s)\}$.

To evaluate the potential efficiency gain of $\widehat{\boldsymbol{\beta}}$ over $\widetilde{\boldsymbol{\beta}}$, we note that the asymptotic variance of $n^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})$ is

$$\boldsymbol{\Sigma}_{\text{IPW}} = \text{var}(\mathbf{F}_{1i}) + \int_0^{\tau} \text{var}(\mathbf{F}_{1i}|T_i^\dagger > s)\frac{S(s)^2 d\Lambda_c(s)}{\pi(s)}$$

It follows that the variance reduction is

$$\begin{aligned}\boldsymbol{\Sigma}_{\text{IPW}} - \boldsymbol{\Sigma}_{t_s} &= \int_0^{t_s} \{\text{var}(\mathbf{F}_{1i}|T_i^\dagger > s) - \text{var}(\mathbf{F}_{2i}+\mathbf{L}_i|T_i^\dagger > s)\}\frac{S(s)^2 d\Lambda_c(s)}{\pi(s)} \\ &\quad + \int_{t_s}^{\tau} \{\text{var}(\mathbf{F}_{1i}|T_i^\dagger > s) - \text{var}(\mathbf{F}_{3i}|T_i^\dagger > s)\}\frac{S(s)^2 d\Lambda_c(s)}{\pi(s)}.\end{aligned}$$

Although it is difficult if not impossible to provide conditions under which $\boldsymbol{\Sigma}_{\text{IPW}} - \boldsymbol{\Sigma}_{t_s}$ is positive definite, we expect the variance of $\widehat{\boldsymbol{\beta}}$ to be smaller than that of $\widetilde{\boldsymbol{\beta}}$ since the $\{Y_{t_s i} - g(\bar{\boldsymbol{\theta}}_{t_s}^{\mathsf{T}}\boldsymbol{\Phi}_i)\} + I(T_i^{\dagger} > t_s)\{Y_{\tau i} - g(\bar{\boldsymbol{\gamma}}_{\tau|t_s}^{\mathsf{T}}\boldsymbol{\Psi}_i)\}$ is expected to have smaller variance than that of $Y_{\tau i} - g(\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i)$ when the τ-GLM is mis-specified and/or $\mathbf{S}_{t_s}$ is highly predictive of Y_τ. To further improve the robustness and efficiency of the proposed procedure, we next detail our final combined estimator that combines information across all $\vec{\mathbf{S}}$ as well as $\widetilde{\boldsymbol{\beta}}$.

S measured at multiple visits When $\mathbf{S}$ is collected over multiple time points, leveraging all measurements to maximally improve estimation efficiency is challenging due to the unknown trade-off between the missing rates and the predictiveness of $\mathbf{S}$ at different time points. While the measurements of $\mathbf{S}$ may be more complete at earlier time points, the latter measurements might be more predictive of Y_τ. We propose to combine all available $\vec{\mathbf{S}}$ by first constructing K estimators, $\widehat{\mathbb{B}} = [\widehat{\boldsymbol{\beta}}_{t_1}, \ ..., \ \widehat{\boldsymbol{\beta}}_{t_K}]_{p\times K}$, with the kth estimator obtained as $\widehat{\boldsymbol{\beta}}$ using $\mathbf{S}_{t_k}$. Using similar arguments as given in Appendix A and B, we may show that $n^{\frac{1}{2}}\{(\widetilde{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})^{\mathsf{T}}, (\widehat{\boldsymbol{\beta}}_{t_1}-\bar{\boldsymbol{\beta}})^{\mathsf{T}}, \ldots, (\widehat{\boldsymbol{\beta}}_{t_K}-\bar{\boldsymbol{\beta}})^{\mathsf{T}}\}^{\mathsf{T}}$ converge jointly to a zero mean multivariate normal. This enables us to construct a combined estimator of $\bar{\boldsymbol{\beta}}$ by deriving an optimal linear combination of $\widetilde{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_{t_1}$, ..., $\widehat{\boldsymbol{\beta}}_{t_K}$. For simplicity, we focus on element-wise combination. For $j = 1, \ldots, p$, we identify the combined estimator

$$\widehat{\beta}_{\text{CMB},j} = \widetilde{\beta}_j - \widehat{\mathbf{W}}_j^{\mathsf{T}} \widehat{\boldsymbol{\Delta}}_j$$

with $\widehat{\mathbf{W}}_j$ being a consistent estimator of

$$\overline{\mathbf{W}}_j = \operatorname*{argmin}_{\mathbf{W}_j} \left\{ \text{var}(\widetilde{\beta}_j - \alpha_j - \mathbf{W}_j^{\mathsf{T}} \widehat{\boldsymbol{\Delta}}_j) \right\},$$

where $\widehat{\boldsymbol{\Delta}}_j = \widetilde{\beta}_j - \widehat{\mathbb{B}}_{j,}$ and for any matrix $\mathbb{B}$, $\mathbb{B}_{j,}$ represents its jth row vector. To obtain $\widehat{\mathbf{W}}_j$ in practice, we approximate the joint distribution of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\mathbb{B}}$ via a perturbation resampling procedure to be detailed in section 2.2. For $b = 1, \ldots, B$, let $\widetilde{\boldsymbol{\beta}}^{(b)}$ and $\widehat{\mathbb{B}}^{(b)}$ denote the bth realization of the resampled estimate of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\mathbb{B}}$, respectively, and let $\widehat{\boldsymbol{\Delta}}_j^{(b)} = \widetilde{\beta}_j^{(b)} - \widehat{\mathbb{B}}_{j,}^{(b)}$. Then we obtain

$$\widehat{\mathbf{W}}_j = \operatorname*{argmin}_{\mathbf{W}_j} \left\{ \sum_{b=1}^{B} \left(\widetilde{\beta}_j^{(b)} - \alpha_j - \mathbf{W}_j^{\mathsf{T}} \widehat{\boldsymbol{\Delta}}_j^{(b)} \right)^2 + \upsilon \|\mathbf{W}_j\|_1 \right\}$$

where $\alpha_j = E(\widetilde{\beta}_j)$ is a nuisance parameter, υ is the tuning parameter and $\|\cdot\|_1$ denote the L_1 norm.

Regularization can be easily adopted to estimate $\boldsymbol{\beta}$ when p is not small relative to the number of events by first noting that $\widetilde{\boldsymbol{\beta}}$ and the proposed augmented estimator $\widehat{\boldsymbol{\beta}}_{t_s}$ are the respective minimizers of $\widetilde{L}_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \widehat{w}_i \ell(Y_{\tau i}, \boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_i)$ and $\widehat{L}_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \ell(\widehat{Y}_{\tau i}^{t_s}, \boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_i)$. To adopt regularization method, such as the adaptive LASSO (Zhang and Lu, 2007), we estimate $\boldsymbol{\beta}$ as $\widetilde{\mathscr{B}}$, the minimizer of the penalized objective function,

$$\widetilde{L}_n(\boldsymbol{\beta}) + \widetilde{\nu}_n \sum_{j=2}^{p} \left|\beta_j/\widetilde{\beta}_j\right| \tag{2.6}$$

where $0 \leq \widetilde{\nu}_n \to \infty$ as and $\widetilde{\nu}_n n^{-\frac{1}{2}} \to 0$ as $n \to \infty$. The regularized counterpart of $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathscr{B}}$, can be obtained as the minimizer of $\widehat{L}_n(\boldsymbol{\beta}) + \widehat{\nu}_n \sum_{j=2}^{p} |\beta_j/\widehat{\beta}_j|$ with similarly chosen $\widehat{\nu}_n$. The resampling procedure as outlined in Section 2.2 can be similarly used to estimate the variability of $\widetilde{\mathscr{B}}$ and $\widehat{\mathscr{B}}$ as well as to construct the final combined estimator that synthesize information on $\mathbf{S}$ across multiple visits.

2.2 Inference via Resampling

To construct $\widehat{\boldsymbol{\beta}}_{\text{CMB}} = (\widehat{\beta}_{\text{CMB},1}, \ldots, \widehat{\boldsymbol{\beta}}_{\text{CMB},p})^{\mathsf{T}}$ and estimate its variance, we propose a perturbation resampling procedure. Specifically, let $\mathbf{V} = (V_1, ..., V_n)^{\mathsf{T}}$ be a vector of independent and identically distributed non-negative random variables with mean 1 and variance 1, generated independent of $\mathscr{D}$. Then for $t_s = t_1, \ldots, t_K$, we obtain a perturbed version of $\widehat{\boldsymbol{\beta}}$ with $\mathbf{Z} = (\mathbf{X}^{\mathsf{T}}, \mathbf{S}_{t_s}^{\mathsf{T}})^{\mathsf{T}}$, $\widehat{\boldsymbol{\beta}}^*_{t_s}$, as the solution to $\widehat{\mathbf{U}}^*_n(\boldsymbol{\beta}) \equiv n^{-1} \sum_{i=1}^{n} V_i \mathbf{X}_i \{\widehat{Y}^*_\tau - g(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_i)\} = 0$, where

$$\widehat{Y}^*_\tau = g(\boldsymbol{\Phi}_i^{\mathsf{T}} \widehat{\boldsymbol{\theta}}^*_{t_s}) + I(T_i > t_s) \widehat{G}^*(t_s)^{-1} g(\boldsymbol{\Psi}_i^{\mathsf{T}} \widehat{\boldsymbol{\gamma}}^*_{\tau|t_s}),$$

$\widehat{\boldsymbol{\theta}}^*_{t_s}$ and $\widehat{\boldsymbol{\gamma}}^*_{\tau|t_s}$ are the respective minimizers of

$$\widehat{\mathbf{Q}}^*_n(\boldsymbol{\theta}) = n^{-1}\sum_{n}^{i=1} \widehat{w}^*_{t_s i}\ell(Y_{t_s i}, \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Phi}_i) + \lambda_1 \mathcal{Q}(|\boldsymbol{\theta}_{[-1]}|),$$

$$\widehat{\mathbf{D}}^*_n(\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^{n} I(T_i > t_s)\widehat{w}^*_{\tau i}\ell(Y_{t_s i}, \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\Psi}_i) + \lambda_2 \mathcal{Q}(|\boldsymbol{\gamma}_{[-1]}|)$$

$\widehat{w}^*_{ti} = \{I(T_i \leq t)\delta_i + I(T_i > t)\}\widehat{G}^*(T_i \wedge t)^{-1}$ and $\widehat{G}^*(\cdot)$ is the weighted Kaplan-Meier estimator of $G(t)$ with $\mathbf{V}$ being the weights. Similarly, we may perturb the IPW estimator $\widetilde{\boldsymbol{\beta}}$ as $\widetilde{\boldsymbol{\beta}}^*$, the solution to the weighted estimating equation

$$\widetilde{\mathbf{U}}^*_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \widehat{w}^*_{\tau i} V_i \mathbf{X}_i\{\widehat{Y}_\tau - g(\boldsymbol{\beta}^\mathsf{T}\mathbf{X}_i)\}.$$

In practice, one can generate B random samples of $\mathbf{V}$ to obtain B realizations of the perturbed estimators $\widetilde{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\beta}}^*_{t_1}, \ldots, \widehat{\boldsymbol{\beta}}^*_{t_K}$. These perturbed estimators can then be used to construct the combined estimator $\widehat{\boldsymbol{\beta}}_{\text{CMB}}$ as described in section 2.1. In addition, it is straightforward to see that the variability in $\widehat{\mathbf{W}}_j$ does not contribute to the variability of $\widehat{\boldsymbol{\beta}}_{\text{CMB}}$ at the first order. Thus, these perturbed samples can also be used to estimate the final variance of $\widehat{\boldsymbol{\beta}}_{\text{CMB}}$ and construct associated confidence intervals.

3 Simulation

We conducted extensive simulation studies to evaluate the finite sample performance of the proposed estimation and inference procedures as well as to compare to existing methods. Throughout, we generated 500 datasets under each configuration at sample size $n = 500$, and $B = 500$ replications were used for the perturbation resampling procedure. For each setting, we obtain the 'true value' $\bar{\boldsymbol{\beta}}$ via monte carlo by averaging over logistic regression estimates obtained from fitting Y_τ against $\mathbf{X}$ using 500 sets of simulated uncensored data at sample size of $N = 10000$. For the proposed estimator, natural spline bases with pre-specified 3 knots for each covariate are used as $\boldsymbol{\Phi}(\cdot)$ and $\boldsymbol{\Psi}(\cdot)$ in the imputation models. In section 3.1, we consider the scenario with $p = 4$ and let $\mathcal{Q}(\cdot) = \|\cdot\|_2$; while in section 3.2, we consider the case with $p = 11$ covariates out of which 7 noise predictors that are unrelated to the risk and let $\mathcal{Q}(\cdot) = \|\cdot\|_1$. For both setting, we let $\tau = 0.8$ and generated a single intermediate outcome S measured at $K = 4$ different time points with $t_1 = 0.05$, $t_2 = 0.1$, $t_3 = 0.15$, $t_4 = 0.2$. The surrogate marker S has an increasing correlation with the outcome over time, but also has increasing proportion of missing values due to censoring or failure. Additional simulation with $\mathbf{S}$ has constant correlation with the outcome over time was also considered. To evaluate the value of S in improving efficiency, we obtained our combination estimator $\widehat{\boldsymbol{\beta}}_{\text{CMB}}$ with $\mathbf{Z} = (\mathbf{X}^\intercal, \mathbf{S}^\intercal)^\intercal$ and with $\mathbf{Z} = \mathbf{X}$, denoted respectively by $\widehat{\boldsymbol{\beta}}_{\text{CMB}}^{\text{KM},\mathbf{Z}}$ and $\widehat{\boldsymbol{\beta}}_{\text{CMB}}^{\text{KM},\mathbf{X}}$. The percent efficiency gain of $\widehat{\boldsymbol{\beta}}$ over $\widetilde{\boldsymbol{\beta}}$ is calculated as $\{\text{MSE}(\widetilde{\boldsymbol{\beta}})/\,\text{MSE}(\widehat{\boldsymbol{\beta}}) - 1\} \times 100$.

In addition to comparing to $\widetilde{\boldsymbol{\beta}}$ (IPW$_{\text{KM}}$, Uno et al., 2007), we also obtained (i) the IPW estimator with censoring weights estimated from fitting a cox model to data $\{(T_i, 1-\delta_i, \mathbf{X}_i), i = 1, ..., n\}$ (IPW$_{\text{Cox},\mathbf{X}}$); (ii) the AIPW estimator (AIPW$_{\text{KM}}$, DiRienzo, 2009) with censoring weights estimated from the Kaplan-Meier and the outcome imputed from the model based on $\Phi(\mathbf{X})$.

3.1 Low dimension setting with $p = 4$

In this setting, we generated $\mathbf{X}_{-1}$, C and $T^{\dagger}$ from

$$\mathbf{X}_{-1} = (X_2, X_3, X_4)^{\mathsf{T}} \sim N(\mathbf{0}, 0.3 + 0.7\mathbb{I}_3), \quad C \sim \text{exponential}(\lambda)$$

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

where $\mathbb{I}_d$ is the $d \times d$ diagonal matrix and $U \sim$ Uniform(0,1). We considered two settings (i) low event rate (12−18% by τ) and heavy censoring rate (65−74% before τ) with $\{\alpha = 12, \lambda = 0.5\}$, where "true" $\boldsymbol{\beta}$ is estimated to be (-1.05,-0.25,-0.13,-0.24); and (ii) moderate event rate (25−34% by τ) and moderate censoring rate (37−50% by τ) with $\{\alpha = 6, \lambda = 1\}$, where "true" $\boldsymbol{\beta}$ is estimated to be (-0.50,-0.27,-0.16,-0.27). We generated $\vec{\mathbf{S}} = (S_{t_1}, S_{t_2}, S_{t_3}, S_{t_4})^{\mathsf{T}}$ from

$$S_t = \text{logit}(U) + 0.1(X_1 + X_2) + (10t^{1.5})^{-1}\varepsilon_t \quad \text{with } \varepsilon_t \sim N(0, 1).$$

where ε_t are generated independently across different time points. Under this setting, the Pearson correlation coefficient between $log(T^{\dagger})$ and $\vec{\mathbf{S}}$ is about $(13\%, 34\%, 55\%, 65\%)^{\intercal}$, while approximately 87%, 76%, 67%, and 60% of patients are at risk at t_1, t_2, t_3 and t_4, respectively. Additional simulation considering the correlation between S_t and outcome is constant over time (about 65%) is used to evaluate how proportion of partially observed subjects impact the efficiency gain.

As shown in Table 1, the proposed estimator has negligible bias and gains substantial efficiency relative to $\mathrm{IPW}_{\mathrm{KM}}$. Compared to $\mathrm{IPW}_{\mathrm{KM}}$, $\mathrm{IPW}_{\mathrm{Cox}}$ and $\mathrm{AIPW}_{\mathrm{KM}}$ attained limited efficiency gain, especially in the low event and high censoring setting. Even in the absence of $\vec{\mathbf{S}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{CMB}}^{\mathrm{KM},\mathbf{X}}$ is much more efficient than $\mathrm{IPW}_{\mathrm{KM}}$ since the imputation model via basis expansion captures the non-linear effects. The proposed estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{CMB}}^{\mathrm{KM},\mathbf{Z}}$ further gains efficiency by additionally incorporating $\vec{\mathbf{S}}$. Figure 1 shows that the efficiency of the proposed estimator $\widehat{\boldsymbol{\beta}}_{t_s}^{\mathrm{KM},\mathbf{Z}}$ relative to $\widetilde{\boldsymbol{\beta}}$ varies substantially across different t_s, and the final estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{CMB}}^{\mathrm{KM},\mathbf{Z}}$, the optimal combination of them, has the highest efficiency gain as expected. The results from Table 1 and Figure 1 also suggest that the proposed interval estimation procedure based on the resampling works well with empirical coverage levels close to the nominal level of 95%. Note that in the setting where S_t has similar correlation with the outcome across different t_s, the efficiency gain of $\widehat{\boldsymbol{\beta}}_{t_s}^{\text{KM},\mathbf{Z}}$ tends to decrease over time (especially in the heavy censoring setting) due to the decreasing proportion of partially observed subjects (i.e., censored between t_s and τ), as shown in Figure 2.

Table 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}	0.58	-2.48	-2.01	-2.39	15.23	20.35	20.50	20.44	0.00	0.00	0.00	
$\text{IPW}_{\text{Cox},\mathbf{x}}$	-1.01	-2.17	-2.26	-2.34	15.17	18.86	18.42	18.62	0.53	16.56	23.17	20.16
AIPW_{KM}	2.01	-2.58	-1.16	-2.32	16.68	22.77	20.00	22.22	-17.75	-20.01	5.69	-15.17
$\text{AUG}_{\text{CMB}}^{\text{KM},\mathbf{x}}$	-0.17	-0.59	-3.41	-0.97	14.76	16.20	14.68	15.29	6.52	59.78	86.67	80.32
$\text{AUG}_{\text{CMB}}^{\text{KM},\mathbf{Z}}$	1.79	0.07	-3.47	-0.22	14.31	14.51	12.95	13.68	11.59	99.51	135.88	126.06
Moderate Event Rate												
IPW_{KM}	0.28	-1.21	-1.06	-0.59	10.87	14.26	13.98	14.19	-	-	-	-
$\text{IPW}_{\text{Cox},\mathbf{x}}$	-0.37	-0.91	-1.07	-0.46	10.78	13.52	12.96	13.41	1.54	11.60	16.25	12.13
AIPW_{KM}	0.73	-1.37	-0.60	-0.63	11.03	13.54	12.11	13.27	-3.21	10.55	33.88	14.28
$\text{AUG}_{\text{CMB}}^{\text{KM},\mathbf{x}}$	-0.50	-0.50	-1.76	0.03	10.61	12.20	11.04	12.09	4.77	37.38	57.34	38.04
$\text{AUG}_{\text{CMB}}^{\text{KM},\mathbf{Z}}$	-0.26	0.44	-1.02	0.33	10.19	11.35	10.34	11.51	13.80	58.77	82.33	52.35

Figure 1: The percent of efficiency gain (%EffG) from the IPW estimator $\widetilde{\boldsymbol{\beta}}$ and the coverage percentage for 95% confidence interval for the proposed estimators $\{\widehat{\boldsymbol{\beta}}_{t_k}^{\mathrm{KM},\mathbf{Z}}, k = 1, 2, 3, 4\}$ as well as the combined estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{CMB}}^{\mathrm{KM},\mathbf{Z}}$ under the low dimension baseline model.

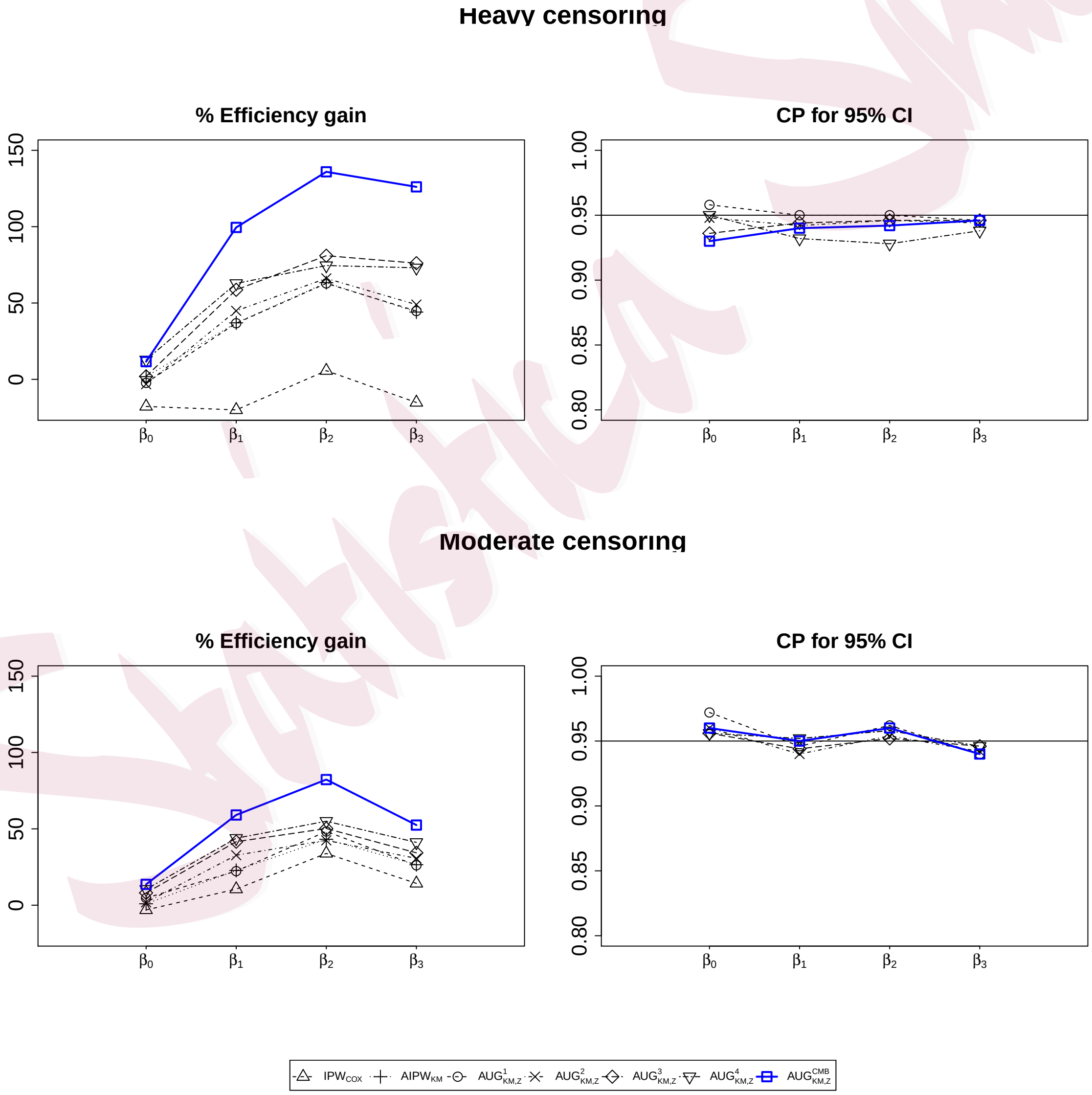
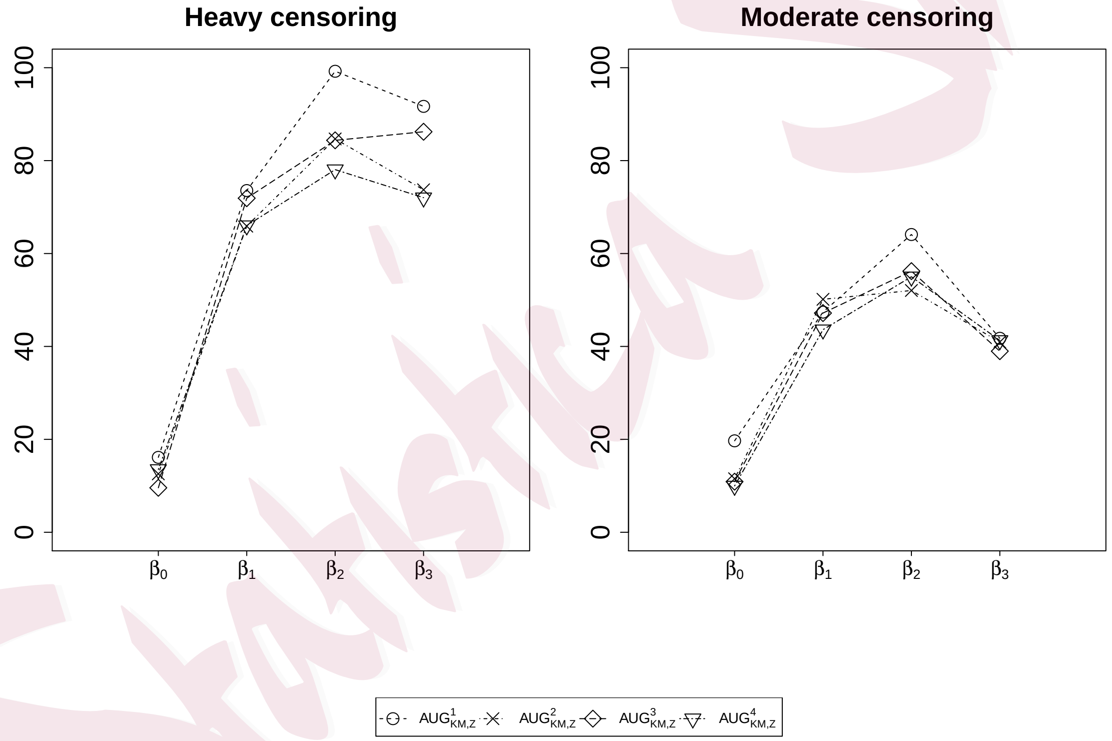


Figure 2: The percent of efficiency gain (%EffG) for the proposed estimators $\{\widehat{\boldsymbol{\beta}}_{t_k}^{\mathrm{KM},\mathbf{Z}}, k = 1, 2, 3, 4\}$ assuming constant correlation between S and outcome over time under the low dimension baseline model.



3.2 Moderate p with regularization

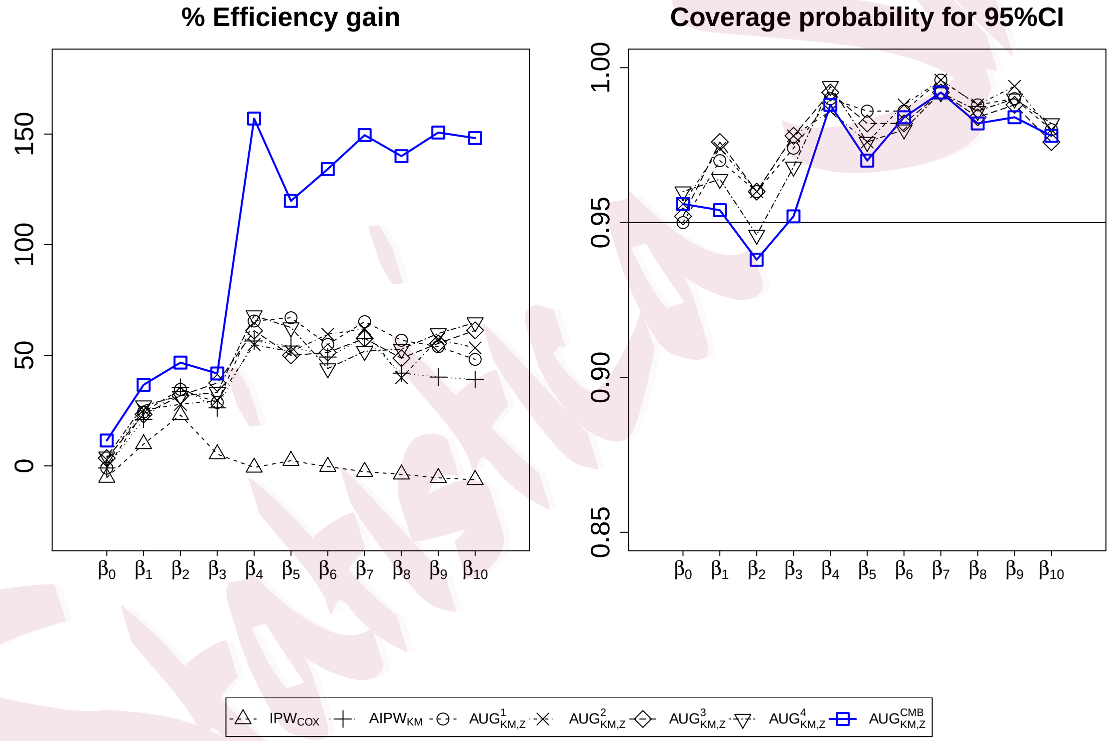
For the setting with $p = 11$, we generated $\mathbf{X}_{-1}$ from independent standard normal and $T^{\dagger}$ from

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

The surrogate markers $\vec{\mathbf{S}}$ was generated the same way as the low dimensional setting and $C \sim$ exponential(1). Under this setting, the observed event rate by τ is about $26 - 38\%$, leading to the effective sample size around 100-200, not large relative to $p = 11$. We use adaptive LASSO as in (2.6) to regularize baseline prediction model in all methods. The "true" $\boldsymbol{\beta}$ for the working model estimated from the complete data is (-0.49, -0.66, -0.52, -0.66, 0.00,0.00, 0.00, 0.00, 0.00, 0.00, 0.00).

Figure 3 summarizes the results for $\{\widehat{\boldsymbol{\beta}}_{t_k}^{\text{KM},\mathbf{Z}}, k = 1, 2, 3, 4\}$, the final combined linear optimal estimator $\widehat{\boldsymbol{\beta}}_{\text{CMB}}^{\text{KM},\mathbf{Z}}$ 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, $\widehat{\boldsymbol{\beta}}_{t_k}^{\text{KM},\mathbf{Z}}$ is more efficient than IPW_{Cox}, and the combined estimator $\widehat{\boldsymbol{\beta}}_{\text{CMB}}^{\text{KM},\mathbf{Z}}$ outperforms all other estimators with substantial efficiency gain over existing methods. The resampling procedures also perform well with empirical coverage percentage ranging from 92-95% for informative signals. The cov-

Figure 3: The percent of efficiency gain (%EffG) and the coverage percentage for 95% confidence interval for the proposed estimator for regularized baseline model



erage percentage for the zero signals range from 96%-98%, which is expected owing to the oracle properties.

4 Example

We illustrate the proposed procedures using a dataset from the Diabetes Prevention Program (DPP) (DPPG, 2002). The DPP is a placebo-controlled randomized clinical trial to investigate whether the change of lifestyle or taking metformin will prevent type 2 diabetes among high-risk adults. The primary outcome, type 2 diabetes, is defined as fasting glucose $\geq$ 140mg/dL for visits through 6/23/1997, or $\geq$ 126 mg/dL for visits on or after 6/24/1997, or 2-h post challenge glucose $\geq$ 200 mg/dL. The study found that the lifestyle intervention, as well as metformin, significantly prevented or delayed the development of type 2 diabetes.

Suppose we are interested in constructing a time-specific risk prediction model for $\tau = 4$ years for the lifestyle intervention group (N=1024) and the placebo group (N=1030), respectively. The event rate was 13.5% for lifestyle intervention group and 27.5% for placebo group by year 4, with 74% and 62% censored before year 4, respectively. The working baseline prediction model includes three predictors: age in ordinal scale, body mass index (BMI) in ordinal scale, and hemoglobin A1c (HBA1C). There are two intermediate outcomes, fasting plasma glucose and HBA1C, which were measured at year 1, 2, and 3.

All covariates are standardized to have mean 0 and standard deviation 1. For the imputation modeling in $\mathrm{AIPW}_{\mathrm{KM}}$ and our approach, we use spline bases with 3 knots for all variables. Resampling with 500 replications is used to generate the variance of the $\mathrm{IPW}_{\mathrm{KM}}$ method and our proposed methods, and bootstrap was used for other methods. As shown in Table 2, the point estimates from $\mathrm{IPW}_{\mathrm{KM}}$ and $\mathrm{IPW}_{\mathrm{Cox}}$ are quite similar, supporting that censoring may be independent of the baseline predictors. The proposed method also provides point estimates similar to $\mathrm{IPW}_{\mathrm{KM}}$ and $\mathrm{IPW}_{\mathrm{Cox}}$, but have substantially smaller standard errors. For example, in the lifestyle intervention group, coefficient estimation for age is -0.21 with standard error 0.15 from $\mathrm{IPW}_{\mathrm{KM}}$ method, while our estimation is -0.25 with standard error 0.095, making the age a significant predictor. Similarly, in the placebo group, coefficient estimation for BMI is 0.13 with standard error 0.138 from $\mathrm{IPW}_{\mathrm{KM}}$ method, while our estimation is 0.15 with standard error 0.062, making the BMI a significant predictor.

Table 2: Estimated prediction models for diabetes by year 3.5 in DPP study

	Coefficient$_{SE}$				Efficiency gain			
	Int	age	BMI	HA1C	Int	age	BMI	HA1C
	Lifestyle group							
IPW$_{\mathrm{KM}}$	$-1.41_{.123}$	$-0.21_{.146}$	$0.15_{.120}$	$0.41_{.146}$	-	-	-	-
IPW$_{\mathrm{Cox},\mathbf{x}}$	$-1.41_{.123}$	$-0.26_{.127}$	$0.21_{.105}$	$0.42_{.148}$	-2.03	32.67	29.61	-2.20
AIPW$_{\mathrm{KM}}$	$-1.28_{.145}$	$-0.17_{.221}$	$0.00_{.202}$	$0.41_{.229}$	-28.25	-56.23	-64.87	-59.08
AUG$^{1}_{\mathrm{KM},\mathbf{x}}$	$-1.36_{.128}$	$-0.23_{.123}$	$0.14_{.096}$	$0.39_{.133}$	-7.66	41.18	55.38	21.46
AUG$^{1}_{\mathrm{KM},\mathbf{Z}}$	$-1.42_{.127}$	$-0.28_{.116}$	$0.10_{.088}$	$0.38_{.122}$	-7.13	58.31	85.66	44.48
AUG$^{2}_{\mathrm{KM},\mathbf{x}}$	$-1.36_{.125}$	$-0.24_{.114}$	$0.17_{.090}$	$0.35_{.128}$	-3.46	64.12	76.26	29.85
AUG$^{2}_{\mathrm{KM},\mathbf{Z}}$	$-1.29_{.133}$	$-0.22_{.116}$	$0.13_{.086}$	$0.34_{.113}$	-15.14	59.29	92.36	69.85
AUG$^{3}_{\mathrm{KM},\mathbf{x}}$	$-1.37_{.123}$	$-0.23_{.101}$	$0.16_{.083}$	$0.31_{.116}$	-1.45	110.49	106.11	59.16
AUG$^{3}_{\mathrm{KM},\mathbf{Z}}$	$-1.43_{.119}$	$-0.22_{.103}$	$0.16_{.083}$	$0.35_{.108}$	5.22	101.59	107.57	85.03
AUG$^{\mathrm{KM},\mathbf{x}}_{\mathrm{CMB}}$	$-1.39_{.121}$	$-0.24_{.099}$	$0.15_{.073}$	$0.29_{.109}$	2.12	119.86	169.57	81.18
AUG$^{\mathrm{KM},\mathbf{Z}}_{\mathrm{CMB}}$	$-1.41_{.119}$	$-0.25_{.095}$	$0.10_{.067}$	$0.31_{.094}$	6.39	137.98	217.79	144.82
	Placebo group							
IPW$_{\mathrm{KM}}$	$-0.58_{.097}$	$0.01_{.134}$	$0.13_{.138}$	$0.34_{.128}$	-	-	-	-
IPW$_{\mathrm{Cox},\mathbf{x}}$	$-0.58_{.098}$	$-0.04_{.120}$	$0.19_{.130}$	$0.37_{.135}$	-1.79	24.24	11.37	-8.98
AIPW$_{\mathrm{KM}}$	$-0.54_{.099}$	$0.09_{.203}$	$0.02_{.210}$	$0.32_{.205}$	-4.25	-56.60	-56.90	-60.74
AUG$^{1}_{\mathrm{KM},\mathbf{x}}$	$-0.58_{.095}$	$-0.02_{.102}$	$0.13_{.113}$	$0.34_{.101}$	4.35	72.02	49.51	60.44
AUG$^{1}_{\mathrm{KM},\mathbf{Z}}$	$-0.54_{.091}$	$0.01_{.091}$	$0.15_{.097}$	$0.39_{.097}$	13.28	117.05	99.78	74.44
AUG$^{2}_{\mathrm{KM},\mathbf{x}}$	$-0.59_{.095}$	$-0.03_{.089}$	$0.15_{.096}$	$0.38_{.093}$	4.34	127.23	105.28	89.12
AUG$^{2}_{\mathrm{KM},\mathbf{Z}}$	$-0.58_{.095}$	$-0.03_{.086}$	$0.15_{.087}$	$0.45_{.093}$	4.72	142.87	147.97	90.12
AUG$^{3}_{\mathrm{KM},\mathbf{x}}$	$-0.59_{.097}$	$-0.04_{.080}$	$0.14_{.081}$	$0.47_{.087}$	0.54	183.43	192.25	118.48
AUG$^{3}_{\mathrm{KM},\mathbf{Z}}$	$-0.63_{.093}$	$-0.03_{.077}$	$0.13_{.076}$	$0.50_{.088}$	9.27	202.67	224.08	137.77
AUG$^{\mathrm{KM},\mathbf{x}}_{\mathrm{CMB}}$	$-0.58_{.095}$	$-0.06_{.070}$	$0.15_{.067}$	$0.45_{.069}$	5.11	267.02	325.85	242.39
AUG$^{\mathrm{KM},\mathbf{Z}}_{\mathrm{CMB}}$	$-0.59_{.088}$	$-0.03_{.065}$	$0.15_{.062}$	$0.45_{.067}$	21.11	327.89	395.28	263.68

5 Remarks

Deriving a robust and efficient estimator for a $\tau-$year risk prediction model is challenging in the presence of heavy censoring prior to τ and potential model mis-specification. Compared to existing literature, there a few key innovations of the proposed approach. First, unlike most existing imputation based estimators, the proposed method is robust to model mis-specifications in both the underlying risk model and the imputation model. Second, our method is able to incorporate information from longitudinal intermediate outcomes that are subject to missingness due to censoring or failure. Third, the proposed efficient data-adaptive combination strategy allows us to effectively combine information from $\mathbf{S}$ measured at different visits along with other consistent estimators (e.g., $\tilde{\boldsymbol{\beta}}$) to achieve maximal efficiency. Analogous to overfitting in regression, our regularization based combination strategy can effectively overcome both the correlation among the estimators and the potentially large number of candidate estimators.

The degree of efficiency gain from incorporating $\mathbf{S}$ in our proposed estimator depends on the censoring distribution prior to τ, how well the τ-GLM approximates the true conditional risk, censoring rate for $\mathbf{S}$, and the predictiveness of $\mathbf{S}$ for Y_τ above and beyond $\mathbf{X}$. The proposed method could be particularly useful in the settings where the prediction model with a long-term outcome involves heavy censoring due to administrative reasons (e.g., study closure, etc.), but intermediate covariates that are predictive of the outcome are collected for the large proportion of patients.

We make the assumption that C is independent of the baseline covariates covariates $\mathbf{X}$ for simplicity. However, similar to existing IPW estimators, we may allow C to depend on $\mathbf{X}$ by calculating the censoring weights $\widehat{w}_i$ by fitting a Cox or other semi-parametric model for $C \mid \mathbf{X}$. When C depends on $\mathbf{X}$ but cannot be correctly modeled, Zhang and Cai (2017) demonstrated via simulation studies that the imputation-based approach tends to be more robust than the simple IPW approach. When p is not small, our approach also has advantages over the augmentation method that uses the Cox model to estimate the censoring weights since employing regularization in the estimation of censoring model would diminish its potential efficiency gain while our imputation based method naturally allows for variable selection.

Throughout, we assume that the intermediate outcomes are potentially measured at the same time points across subjects. This is a reasonable assumption for clinical trials since study visit times are typically prescheduled according to the study protocol. For setting where ascertainment times vary across patients, we can choose a set of $\{t_s, s = 1..., K\}$ as landmark time points and summarize **S** information up to t_s as the intermediate outcome associated with t_s for each patient.

Supplementary Materials

The online supplementary materials include the appendix for technical details of the proof.

Acknowledgements

The authors thank Diabetes Prevention Program for providing the data to use as an example. This research was partially supported by NIH grants T32AI007358, R01HL089778, and R01GM085047.

References

Bang, H. and Robins, J. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models *Biometrics 61(4)*, pp. 962–973.

Cai, T and Cheng, S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times *Biostatistics 9(2)*, pp. 216–233.

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

DiRienzo, G. (2009). Flexible Regression Model Selection for Survival probabilities: with Application to AIDS. *Biometrics 65*, pp. 1194–1202.

Diabetes Prevention Program Research Group (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine 346*, pp. 393-403.

Eguchi, S. and Copas, J. (2002). A class of logistic-type discriminant functions *Biometrika* , pp. 1-22.

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

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

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

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

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

Van Houwelingen, Hans C (2007). *Dynamic Predicitonin Clinical Survival Analysis*. *Scandinavian Journal of Statistics 34(1)*, pp. 70–85.

Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika 94(3)*, pp. 691–703.

Zheng, Y. and Cai, T. (2017). Augmented Estimation for t-year Survival with Censored Regression Model. *Biometrics 73(4)*, pp. 1169-1178.

First author: Yu Zheng

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

E-mail: (ezheng@sdac.harvard.edu)

Second author: Lu Tian

Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305

E-mail: (lutian@stanford.edu)

Third author: Tianxi Cai

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

E-mail: (tcai@hsph.harvard.edu)