Title

Estimating Treatment Effect under Additive Hazards Models with High-dimensional Covariates

Authors

Hou, Jue
Bradic, Jelena
Xu, Ronghui

Publication Date

2019-06-29

Peer reviewed

Estimating Treatment Effect under Additive Hazards Models with High-dimensional Covariates

Jue Hou[1], Jelena Bradic[1], and Ronghui Xu[1,2]

[1]Department of Mathematics, University of California San Diego
[2]Department of Family Medicine and Public Health, University of California San Diego

Abstract

Estimating causal effects for survival outcomes in the high-dimensional setting is an extremely important topic for many biomedical applications as well as areas of social sciences. We propose a new orthogonal score method for treatment effect estimation and inference that results in asymptotically valid confidence intervals assuming only good estimation properties of the hazard outcome model and the conditional probability of treatment. This guarantee allows us to provide valid inference for the conditional treatment effect under the high-dimensional additive hazards model under considerably more generality than existing approaches. In addition, we develop a new Hazards Difference (HDi) estimator. We showcase that our approach has double-robustness properties in high dimensions: with cross-fitting the HDi estimate is consistent under a wide variety of treatment assignment models; the HDi estimate is also consistent when the hazards model is misspecified and instead the true data generating mechanism follows a partially linear additive hazards model. We further develop a novel sparsity doubly robust result, where either the outcome or the treatment model can be a fully dense high-dimensional model. We apply our methods to study the treatment effect of radical prostatectomy versus conservative management for prostate cancer patients using the SEER-Medicare Linked Data.

Keywords: binary treatment; confounding; double robustness; orthogonal score; survival outcome.

1 Introduction

Treatment effect estimation and inference is an essential topic of interest in causal inference and causal discoveries. It has drawn tremendous interest spanning many different fields. Our work was motivated by the proliferation of 'big, observational data' from Electronic Medical/Health Records (EMR/EHR), which provides an abundant resource for studying the effect of various treatments and serves as the alternative or exploratory when a randomized trial is implausible or uneconomical. With the availability of such large databases the challenge in studying causal effects, is to handle a large "$p \gg n$" number of potential confounders. Motived by studies in cancer, using the linked Surveillance, Epidemiology, and End Results (SEER) - Medicare database, our primary focus are causal effects on a survival outcome.

Traditionally, high-dimensional analysis with right censored data have largely stopped at performing variable selection only: inference is then reported on the findings with only those selected covariates. However, such findings can be spurious due to the many difficulties in selecting "only" the correct features; wrong feature selection can adversely affect the causal discoveries especially so in the observational data setting. We seek to address this challenge by developing a powerful, new method for estimating treatment effects that yields valid asymptotic confidence intervals for the treatment effects under the additive hazards models, in the presence of possible wrong selections.

We focus on a family of orthogonal scores, introduced in as early as Neyman (1959). To the best of our knowledge, we develop the first such score for the additive hazards by decoupling the at-risk process with treatment assignment. We develop an orthogonal-score-based method for treatment effect estimation that allows for valid statistical inference and tractable asymptotic theory, with low or high dimensional covariates. Asymptotic normality results are especially important in the causal inference setting. Yet, asymptotics in high-dimensional additive hazards models have been largely left open. This paper addresses these limitations. We then proceed to develop a new difference in hazards estimate (HDi) that utilizes covariate balancing and is a special case of the orthogonal score family. We showcase a double-robustness property of this estimate, which seems to be unique in both low and high dimensional setting, and is therefore of independent interest. We conclude by establishing consistency of estimation under model misspecification, both in terms of sparsity and the classical model misspecification in terms of their functional forms. To better characterize the asymptotic behavior of the estimator without sparsity, we move away from concentration at the true or "least false" parameter (Hjort, 1986, 1992), and build our arguments around a new magnitude structure and simple and intuitive cross-validation.

1.1 Related work

The first main technical contribution of our work is an asymptotic normality theory enabling statistical inference in high-dimensional models for right censored survival data. Recent results of Bradic et al. (2011); Hou et al. (2019); Yu et al. (2019) provide asymptotic properties under the high dimensional Cox type proportional hazards models, which include competing risks. To the best of our knowledge, however, we provide the first results on high-dimensional inference including confidence interval construction under the additive hazards models.

Orthogonal scores, despite being a long familiar concept in the semiparametric literature (Bickel et al., 1998; Newey, 1990), have not been explored for additive hazards models up to date. Orthogonal scores relate to the profile likelihood as well as the least favorable direction in likelihood inference; including efficient score drawn from the profile likelihood (special case of the orthogonal score) and nonparametric likelihood for semiparametric models, for example (Murphy and van der Vaart, 2000; Severini and Wong, 1992). However, nonparametric likelihood is not applicable for the additive hazards model (Lin and Ying, 1994; Martinussen and Scheike, 2007). Moreover, efficient scores have been derived separately under

this model. However, regretably they typically require the knowledge (or independent estimation) of the baseline hazard function (Dukes et al., 2019; Lin and Ying, 1994).

Benefits of the orthogonal score have long been known: an estimator obtained from such score should not be affected by the slower than root-n convergence in the estimation of the nuisance parameters in the model (Bickel et al., 1998; Newey, 1990). This property was recently utilized for purposes of estimating treatment effects in high dimensional models (Belloni et al., 2013; Chernozhukov et al., 2018a; Farrell, 2015). However, models with censored observations present a considerable challenge. Approaches based on uncensored data do not automatically generalize to censored data; complex dependencies are induced by the presence of censoring.

There is a connection between orthogonal score and double robustness that has not always been made explicit in the literature. For missing data in general there can be two working models, one for the outcome and another for the treatment assignment. An estimator is doubly robust (DR) if it is consistent, as long as one (but not necessarily both) of the two working models is correct (Bang and Robins, 2005; Robins and Rotnitzky, 1995, 2001). In high dimensions Farrell (2015) showed that the DR estimator in Robins and Rotnitzky (1995) is still consistent when $p \gg n$. On the other hand, DR estimators for censored outcomes have only been considered in low-dimensional setting, with a fixed $p < n$. In the context of additive hazards model, Wang et al. (2017) considered DR estimators constructed through IPW approaches. Kang et al. (2018) extended the DR approach of Robins et al. (1992) but relied on kernel density estimators which are not suitable for high-dimensional covariates. The recently published work of Dukes et al. (2019) is closely related to ours, and will be discussed in more details later; but again, their work is under the low-dimensional setting.

Finally we note a growing literature on treatment effect estimation that makes use of different machine-learning methods (Athey and Imbens, 2016; Chernozhukov et al., 2018a,b; Farrell et al., 2018; Imai and Ratkovic, 2013; Powers et al., 2018; Wager and Athey, 2018). None of the above, however, deals with censored survival data, which is often encountered in analyzing EMR/EHR data, as in our motivating application.

1.2 Organization

The organization of the rest of the paper is as follows. In Section 2, we propose a family of inferential methods based on orthogonal scores for the treatment effect under the additive hazards model for the survival outcome and the logistic regression model for treatment assignment. In Section 3, we develop doubly robust estimation, a hazard difference (HDi) estimator, cross-fitting composite estimation with cross-validation results. Section 4 contains extensive simulation studies illustrating stable properties of the newly proposed estimates across a number of settings in high-dimensions. In Section 5, we apply our methods to the study of treatment effect of surgery, radical prostatectomy, versus conservative disease management, using the SEER-Medicare Linked Data. Section 6 contains the conclusion and discussion. The detailed proofs of the theoretical results are given in the Supplementary Materials.

2 Orthogonal score and inference for treatment effect

2.1 Orthogonal Score

We consider right censored observations where T and C denote the event and the censoring time, respectively. Our observations consist of n independent and identically distributed (i.i.d.) samples $W_i = (X_i, \delta_i, D_i, Z_i)$, where $X_i = \min\{T_i, C_i\}$, $\delta_i = I(T_i \leq C_i)$, $D_i \in \{0.1\}$ is the treatment assignment and

$Z_i \in \mathbb{R}^p$ denotes a $p \times 1$ vector of covariates. Let $\lambda(t; D, Z)$ be the conditional hazard function of T given D and Z. We consider finite study duration and denote the upper limit of follow-up time as $\tau < \infty$. The additive hazards model for $\lambda(t; D, Z)$ is

$$\lambda(t; D, Z) = \lambda_0(t) + D\theta + \beta^\top Z, \tag{1}$$

where $\lambda_0(\cdot)$ is an unknown baseline hazard function. Following the convention of Andersen and Gill (1982), we denote the counting process and at-risk process for subjects $i = 1, \ldots, n$ with $N_i(t) = \delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$, respectively. We can rewrite the additive hazards model (1) as

$$dN_i(t) = \{Y_i(t)D_i dt\}\theta + \beta^\top Z_i Y_i(t)dt + Y_i(t)d\Lambda(t) + dM_i(t; \theta, \beta, \Lambda), \tag{2}$$

where $\Lambda(t)$ is the baseline cumulative hazard and $M_i(t; \theta, \beta, \Lambda)$ is a martingale with respect to the filtration $\mathcal{F}_{n,t} = \sigma\{N_i(u), Y_i(u), D_i, Z_i : u \leq t, i = 1, \ldots, n\}$ when evaluated at the true parameter values. Our goal is to draw inference for the treatment effect, θ, while allowing the dimension of the covariates, p, to be much larger than the sample size n.

In order to develop an orthogonal score for θ under the additive hazards model (1), we make use of a model for the conditional probability of treatment assignment, often called the propensity score in the causal inference literature. In this case we assume a logistic regression model:

$$\mathbb{P}(D = 1|Z) = \exp(\gamma^\top \tilde{Z})/(1 + \exp(\gamma^\top \tilde{Z})) := \text{expit}(\gamma^\top \tilde{Z}), \tag{3}$$

where $\tilde{Z} = (1, Z_1, \ldots, Z_p)^\top$ represents the vector of covariates with the intercept term.

Under models (1) and (3), writing the nuisance parameter $\eta = (\beta, \Lambda, \gamma)$, a score function, hereby denoted by $\psi(\theta, \eta)$, is defined to be an orthogonal score for θ, if the Gâteaux derivative with respect to η

$$\frac{\partial}{\partial r}\mathbb{E}\{\psi(\theta_0; \eta_0 + r\triangle\eta)\}\Big|_{r=0} = 0, \tag{4}$$

where θ_0 and η_0 are the true values, respectively, and $\triangle\eta = \eta - \eta_0$. In other words, the orthogonality of a score function is defined as the local invariance of the score to a small perturbation in the nuisance parameter around the true parameters. Under orthogonality, the estimation of the treatment effect is not affected by the convergence rate of any consistent estimation of the nuisance parameter (Newey, 1994). Because of this, orthogonal scores can be very useful in high dimensional inference (Chernozhukov et al., 2018a).

Our construction of an orthogonal score is inspired by the efficient scores of Robins and Rotnitzky (1995) and Hahn (1998), and a closely related score of Robins (2004). A common approach for these scores is based on a product of two residuals: one from the outcome model and one from the treatment model. Under our models a natural candidate would be the product of the martingale residuals and the logistic regression residuals. However, it is not difficult to see that such a product would not be orthogonal according to the definition (4), ultimately due to the dependence between the at-risk process and the treatment assignment. We discover that, if we are willing to assume:

$$C \perp\!\!\!\perp (T, D)|Z, \tag{5}$$

where '$\perp\!\!\!\perp$' denotes statistical independence, we can make a simple correction, by 'decoupling' the at-risk process and the treatment assignment. The resulting orthogonal score is given in the lemma below.

Lemma 1. *Under models (1) and (3) and assumption (5), the score*

$$\phi(\theta;\beta,\Lambda,\gamma)=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\exp\left(D_i\theta t\right)\left(D_i-\operatorname{expit}(\gamma^{\top}\tilde{Z}_i)\right)dM_i(t;\theta,\beta,\Lambda) \tag{6}$$

identifies the true parameters $(\theta_0;\beta_0,\Lambda_0,\gamma_0)$, i.e., $\mathbb{E}[\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)]=0$. Moreover, ϕ is an orthogonal score for θ in the sense of (4).

Remark 1. To utilize the orthogonal score to estimate θ, (6) is seen as an equation for θ only. Plugging in consistent estimates of the nuisance parameters (β,Λ,γ), (6) is solved for θ only to obtain $\hat{\theta}$.

We note that (6) is a special case of the class of estimating equations considered recently in an independent work of Dukes et al. (2019). Therein the authors also work under the Condition (5). While (5) is stronger than the usual noninformative censoring, $C \perp\!\!\!\perp T|(D,Z)$, however, observe that under the two models (1) and (3), D also plays the role of a 'response' in relation to Z. When considering these two models simultaneously, it is perhaps natural to a certain extent to require that the censoring mechanism be independent of D given Z. Nevertheless, under the usual condition, $C \perp\!\!\!\perp T|(D,Z)$, we can also decouple $Y_i(t)$ and D to similarly construct an orthogonal score, as stated in the following lemma.

Lemma 2. *Assume $C \perp\!\!\!\perp T|(D,Z)$ and models (1) and (3), the score*

$$\phi_C(\theta;\beta,\Lambda,\gamma,S_C)=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\exp\left(D_i\theta t\right)S_C^{-1}(t|D_i,Z_i)\left(D_i-\operatorname{expit}(\gamma^{\top}\tilde{Z}_i)\right)dM_i(t;\theta,\beta,\Lambda),$$

identifies the true parameters $(\theta_0;\beta_0,\Lambda_0,\gamma_0,S_{C,0})$, where $S_{C,0}$ is the true value of $S_C(t|D,Z)=\mathbb{P}(C_i\geq t|D_i,Z_i)$, and ϕ_C is an orthogonal score for θ.

Remark 2. To actually use Lemma 2 to estimate θ, one needs to obtain an estimate of $S_C(t|D,Z)=\mathbb{P}(C_i\geq t|D_i,Z_i)$, which in the presence of high dimensional covariates would be close to impossible (effective density estimation beyond dimension 10 or arguably 5 is not developed).

For the rest of this paper we assume (5). Note that (5) is easily satisfied in case of administrative censoring, i.e. caused by the end of a study.

2.2 From orthogonality to double robustness

In this subsection, we highlight the fact that the proposed orthogonal score has the usual doubly robust property in terms of misspecification of either model (1) or (3).

Lemma 3. *(a) Suppose that model (1) is true, while the treatment assignment D follows a nonparametric model*

$$\mathbb{P}(D=1|Z)=m_0(Z). \tag{7}$$

For any given γ^*, $\theta=\theta_0$ is the root of the equation

$$\mathbb{E}[\phi(\theta;\beta_0,\Lambda_0,\gamma^*)]=0.$$

(b) Suppose that model (3) is true, while T follows a partially linear additive hazards model

$$\lambda(t,D,Z)=D\theta+g_0(t;Z), \tag{8}$$

with g_0 being unspecified function of both time and covariates. Under Condition (5), for any given β^* and Λ^* with bounded total variation, $\theta=\theta_0$ is the root of the equation

$$\mathbb{E}[\phi(\theta;\beta^*,\Lambda^*,\gamma_0)]=0.$$

Lemma 3 implies that we can use our score to produce estimator of the treatment effect that is doubly robust to model misspecification in the classical low dimensional sense. In Section 3, we develop the approach for high dimensional double robustness which is quite different.

2.3 One-shot inference for treatment effect

In this subsection we present asymptotic results named one-shot inference for treatment effect, where all of the data, rather than subsamples, are utilized twice: first to estimate the unknown nuisance parameters and then to solve the score equation below, defining $\hat{\theta}$ as the solution to

$$\phi(\theta; \hat{\beta}, \hat{\Lambda}, \hat{\gamma}) = 0.$$

In the above, $\hat{\beta}, \hat{\Lambda}, \hat{\gamma}$ are estimators of the unknown nuisance parameters. We don't restrict the particular choice of these estimators, as long as they satisfy the conditions specified below. We do note, however, that $\hat{\Lambda}$ is typically obtained by a Breslow type estimator (Lin and Ying, 1994) which depends on the values of $(\hat{\beta}, \theta)$; here we may allow θ to remain unknown for later purposes, and abbreviate $\hat{\Lambda}(t; \hat{\beta}, \theta)$ as simply $\hat{\Lambda}(t; \theta)$.

Before stating the assumptions, we define the *average training deviance* corresponding to model (1) and (3) as, respectively,

$$\mathcal{D}^2_\beta(\hat{\beta}, \beta_0) = n^{-1} \sum_{i=1}^{n} \int_0^\tau \left\{ (\hat{\beta} - \beta_0)^\top Z_i \right\}^2 Y_i(t) dt,$$
$$\mathcal{D}^2_\gamma(\hat{\gamma}, \gamma_0) = n^{-1} \sum_{i=1}^{n} \left\{ \text{expit}\left(\hat{\gamma}^\top Z_i\right) - \text{expit}\left(\gamma_0^\top Z_i\right) \right\}^2.$$

Note that $\mathcal{D}^2_\beta(\hat{\beta}, \beta_0)$ is the same as the Bregman divergence used in Gaïffas and Guilloux (2012), and $\mathcal{D}^2_\gamma(\hat{\gamma}, \gamma_0)$ is the excess risk as in van de Geer (2008).

Assumption 1.

(i) The survival outcome follows model (1) whereas the treatment follows model (3);

(ii) $C \perp\!\!\!\perp (T, D) | Z$;

(iii) $\mathbb{P}\left(\sup_{i=1,\ldots,n} \|Z_i\|_\infty < K_Z\right) = 1$;

(iv) the baseline hazard is bounded, $\sup_{t \in [0,\tau]} \lambda_0(t) < K_\Lambda$;

(v) positive at-risk rate on the overlap set $\mathbb{E}\{\mathbb{E}(Y(\tau)|Z; D = 0)\text{Var}(D|Z)\} \geq \varepsilon_Y > 0$;

(vi) positive event rate on the overlap set $\mathbb{E}\{\mathbb{E}(N(\tau)|Z; D = 0)\text{Var}(D|Z)\} \geq \varepsilon_N > 0$;

(vii) the total variation of $\hat{\Lambda}(\cdot; \theta)$ is bounded by K_v uniformly in θ with probability tending to one as $n \to \infty$;

(viii) $\hat{\Lambda}(t; \theta)$ is approximately linear in θ in the neighborhood of θ_0, with respect to the total variation; with arbitrary partition $0 = t_0 < \cdots < t_N = \tau$ and $N \in \mathbb{N}$,

$$\bigvee_{t=0}^{\tau} \left\{ \hat{\Lambda}(t; \theta) - \hat{\Lambda}(t; \theta_0) \right\} = \sup_{\substack{0 = t_0 < \cdots < t_N = \tau \\ N \in \mathbb{N}}} \sum_{j=1}^{N} \left| \hat{\Lambda}(t_{j-1}; \theta) - \hat{\Lambda}(t_j; \theta_0) \right| = O_p(|\theta - \theta_0|);$$

(ix) the rates of estimation errors satisfy

$$\sqrt{\log(p)}\|\hat{\beta}-\beta_0\|_1+\sup_{t\in[0,\tau]}|\hat{\Lambda}(t;\theta_0)-\Lambda_0(t)|+\|\hat{\gamma}-\gamma_0\|_1$$
$$+\sqrt{n}\mathcal{D}_\gamma(\hat{\gamma},\gamma_0)\left(\mathcal{D}_\beta(\hat{\beta},\beta_0)+\sup_{t\in[0,\tau]}|\hat{\Lambda}(t;\theta_0)-\Lambda_0(t)|\right)=o_p(1), \tag{9}$$

and the estimation error of the baseline hazard satisfies additionally

$$\int_0^\tau H(t)d\{\hat{\Lambda}(t,\theta_0)-\Lambda_0(t)\}=o_p(1) \tag{10}$$

for any process $H(t)$ with $\sup_{t\in[0,\tau]}|H(t)|=O_p(1)$ and adapted to the filtration $\mathcal{F}_{n,t}=\sigma\{N_i(u),Y_i(u),D_i,Z_i: u\le t, i=1,\ldots,n\}$.

Under Assumption 1 above, we have the following result regarding the asymptotic distribution of $\hat{\theta}$ in the presence of both high-dimensional and infinite-dimensional nuisance parameters. The result below does not require consistent model selection for either the outcome or the treatment model, and is based solely on good prediction properties.

Theorem 1. *Under Assumption 1, $\hat{\theta}$ that solves $\phi(\theta;\hat{\beta},\hat{\Lambda}(\theta),\hat{\gamma})=0$, converges in distribution to a normal random variable at $\sqrt{n}$-rate,*

$$\hat{\sigma}^{-1}\sqrt{n}(\hat{\theta}-\theta_0)\rightsquigarrow N(0,1),$$

where the variance estimator takes the closed form

$$\hat{\sigma}^2=\frac{n^{-1}\sum_{i=1}^n\delta_i\{D_i-\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)\}^2e^{2\hat{\theta}D_iX_i}}{\left\{n^{-1}\sum_{i=1}^n(1-D_i)\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)X_i\right\}^2}.$$

A few comments are in order. Bounded norm of the covariates as well as bounded baseline hazard, Assumptions 1-iii and 1-iv, appear commonly in non-linear high dimensional models including the generalized linear model (van de Geer et al., 2014) and the Cox proportional hazards regression model (Huang et al., 2013), as well as the additive hazards models (Lin and Lv, 2013). Assumptions 1-v and 1-vi require the at-risk as well as the event rate to be bounded above zero on some overlap set in the covariate space, i.e. where $\text{Var}(D|Z)$ or equivalently, the probabilities of assignment to treatment and control, is bounded above zero. We acknowledge that such assumptions are made possible by the regression models that are postulated over the whole covariate space, and care needs to be taken in interpretation of the results in practice if positivity assumption does not hold (Westreich and Cole, 2010).

The conditions in Assumption 1 collectively play a similar role as the commonly used, high-level requirement in semiparametrics where each estimator is required to converge at $n^{-1/4}$ rate or faster (Belloni et al., 2013, 2015; Chernozhukov et al., 2018a; Farrell, 2015; Robinson, 1988). Here, our conditions are weaker, allowing one or two (but not all simultaneously) of our estimators to converge arbitrarily slow as long as the products

$$\|\hat{\gamma}-\gamma_0\|_2\|\hat{\beta}-\beta_0\|_2 \qquad \text{and} \qquad \|\hat{\gamma}-\gamma_0\|_2\sup_{t\in[0,\tau]}|\hat{\Lambda}(t;\theta_0)-\Lambda_0(t)|$$

converge at the $n^{-1/2}$ rate. The latter product is specific for the semiparametric additive hazards model and is new in the literature, namely, the interference of the infinite-dimensional Λ and the treatment model coefficients.

The proof of Theorem 1 has two parts. In the first part, we establish the asymptotic equivalence

$$\phi(\theta; \hat{\beta}, \hat{\Lambda}(\cdot, \theta), \hat{\gamma}) = \phi(\theta; \beta_0, \Lambda_0, \gamma_0) + o_p\left(\sqrt{n}|\theta - \theta_0| + 1\right),$$

by utilizing that fact that the orthogonality score is insensitive to perturbations in the nuisance parameters. In the second part, we use the identifiability from Lemma 1 to establish the asymptotic normality of $\hat{\theta}$. Our proof technique allows $\beta_0^\top Z_i$ to grow arbitrarily large, as $\|\beta_0\|_1$ grows with p and n. Our asymptotic normality is the first result established with unbounded hazards of a survival outcome, which distinguishes us from the existing literature (Hou et al., 2019; Yu et al., 2019).

It is worth pointing out that the above results hold if we replace models (3) and (1) with the nonparametric (7) and the partially additive model (8), respectively. The rate conditions 1-ix would then take the form

$$\sup_z |\hat{m}(z) - m_0(z)| = o_p(1), \quad \sup_{z, t\in[0,\tau]} |\hat{g}(t; z) - g_0(t; z)| = o_p(1)$$

$$\text{and } \frac{1}{n}\sum_{i=1}^n \{\hat{m}(Z_i) - m_0(Z_i)\}^2 \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\hat{g}(t; Z_i) - g_0(t; Z_i)\}^2 Y_i(t)dt = o_p(n^{-1}).$$

With suitable machine learning methods for $\hat{m}(z)$ and $\hat{g}(t; z)$, we may draw inference on the treatment effect θ if $\hat{m}(z)$ and $\hat{g}(t; z)$ are uniformly consistent and the product of their mean squared training errors is of the order n^{-1}. In Section 3.2 we will use cross-fitting to relax the uniform consistency requirement.

2.4 An illustrative case: Lasso regularization

Whenever models (1) and (3) are sparse high-dimensional models, many regularization methods can be employed for obtaining $\hat{\beta}$ and $\hat{\gamma}$. We provide illustration here for a simple case, of strictly sparse models with $s_\beta = \|\beta_0\|_0$, $s_\gamma = \|\gamma_0\|_0$, and regularization by Lasso. However, note that our results apply to many different regularization methods that correspond to different sparsity patterns (e.g., group, hierarchical, non-convex or ridge type of regularization).

For the additive hazards survival outcome model (1), a simple and widely used approach is the Lasso regularized estimator of Leng and Ma (2007), defined as

$$(\hat{\theta}_l, \hat{\beta}^\top)^\top = \underset{(\theta_l, \beta^\top)^\top \in \mathbb{R}^{p+1}}{\operatorname{argmin}} (\theta_l, \beta^\top) H_n (\theta_l, \beta^\top)^\top - 2(\theta_l, \beta^\top) h_n + \lambda(\|\beta\|_1 + |\theta_l|), \tag{11}$$

where

$$\begin{aligned} H_n =& n^{-1} \sum_{i=1}^n \int_0^\tau \left\{ \left(D, Z_i^\top\right)^\top - \left(\bar{D}(t), \bar{Z}(t)^\top\right)^\top \right\}^{\otimes 2} Y_i(t)dt, \\ h_n =& n^{-1} \sum_{i=1}^n \int_0^\tau \left\{ \left(D, Z_i^\top\right)^\top - \left(\bar{D}(t), \bar{Z}(t)^\top\right)^\top \right\} dN_i(t), \end{aligned}$$

with $\bar{D}(t) = \sum_{i=1}^n D_i Y_i(t) / \sum_{i=1}^n Y_i(t)$ and $\bar{Z}(t) = \sum_{i=1}^n Z_i Y_i(t) / \sum_{i=1}^n Y_i(t)$.

The estimation of γ under model (3) is similar. We may use the Lasso estimator under the logistic regression model (Shevade and Keerthi, 2003):

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^{p+1}}{\operatorname{argmin}} -\frac{1}{n}\sum_{i=1}^n \left\{ \{D_i \gamma^\top \tilde{Z}_i - \log(1 + \exp\{\gamma^\top \tilde{Z}_i\}) \right\} + \lambda\|\gamma\|_1. \tag{12}$$

We used the same notation λ for the tuning parameter in (11) and (12) for simplicity only.

Prediction properties of the above regularized estimators have been established under either the restricted eigenvalue condition (Bickel et al., 2009) or the compatibility condition (van de Geer and Bühlmann, 2009; van de Geer, 2007). Following Gaïffas and Guilloux (2012), Zhang et al. (2017) and van de Geer (2008), under regularity conditions that are weaker than those needed for tight model selection, the above estimators satisfy

$$\|\hat{\beta}-\beta_0\|_1 = O_p\left(s_\beta\sqrt{\log(p)/n}\right),\ \|\hat{\gamma}-\gamma_0\|_1 = O_p\left(s_\gamma\sqrt{\log(p)/n}\right),$$
$$\mathcal{D}_\beta(\hat{\beta},\beta_0) = O_p\left(\sqrt{s_\beta\log(p)/n}\right),\ \mathcal{D}_\gamma(\hat{\gamma},\gamma_0) = O_p\left(\sqrt{s_\gamma\log(p)/n}\right),$$

whenever $\lambda > C\sqrt{\log(p)/n}$ and $\log(p)/n = o(1)$, i.e. allowing the dimension p to be much larger than n.

Remark 3. From the above we can see that a sufficient condition for Theorem 1 is on the sparsities of outcome and treatment models:

$$s_\beta = o(\sqrt{n}/\log(p)),\ s_\gamma = o(\sqrt{n}/\log(p)),\ \text{and}\ s_\beta s_\gamma = o\left(n/\log(p)\right),$$

provided that the Lasso estimators are used.

Our 'one-shot' estimator achieves the rate condition comparable to those of Belloni et al. (2013), Farrell (2015) and Tan (2018), and slightly better than that of Belloni et al. (2015).

Finally, the traditional Breslow estimator of the baseline cumulative hazard is $\hat{\Lambda}(t;\hat{\beta},\hat{\theta}_l)$ (Lin and Ying, 1994), where

$$\int_0^t \frac{\sum_{i=1}^n\{dN_i(u)-Y_i(u)(\hat{\beta}^\top Z_i+\theta D_i)du\}}{\sum_{i=1}^n Y_i(u)} := \hat{\Lambda}(t;\hat{\beta},\theta). \tag{13}$$

It has a rate of uniform convergence $O_p\left(\min\left\{\|\hat{\beta}-\beta_0\|_1, \mathcal{D}_\beta(\hat{\beta},\beta_0)\right\}+|\hat{\theta}_l-\theta_0|\right)$ under Conditions 1-i, 1-iii, 1-iv and 1-v, following the expansion

$$\hat{\Lambda}(t;\hat{\beta},\hat{\theta}_l) = \Lambda_0(t) + \int_0^t \frac{\sum_{i=1}^n dM_i(u;\theta_0,\beta_0,\Lambda_0)}{\sum_{i=1}^n Y_i(u)} - \int_0^t\{(\hat{\beta}-\beta_0)^\top\bar{Z}(u)+(\hat{\theta}_l-\theta_0)\bar{D}(u)\}du. \tag{14}$$

As mentioned earlier, Theorem 1 applies to both $\hat{\Lambda}(t;\hat{\beta},\hat{\theta}_l)$ and $\hat{\Lambda}(t;\hat{\beta},\theta)$; the latter is particularly useful for double robustness properties in Section 3 below.

3 Exploring Double Robustness

Double robustness in high-dimensional settings attracted a lot of attention recently. Different from the low-dimensional settings where the definition of DR takes one commonly acknowledged meaning, in high dimensions multiple interpretations exist. Our focus in this section will be on the usual model misspecification as well as the sparsity relaxation; the latter focuses on relaxing sparsity assumptions and allowing possible fully dense ultra high dimensional models.

When we allow fully dense but well specified models, consistent estimation of all the model parameters including the nuisance ones is not possible. DR estimates that relax sparsity assumption have been developed for a handful of specific models; both Chernozhukov et al. (2018b) and Bradic et al. (2018) can be applied to linear outcome and linear treatment models only. In contrast, we consider additive hazards outcome model and logistic treatment model - a setting required by applications like ours, as

well as presenting substantial theoretical and methodological challenges. Throughout our work, unknown censoring distribution is always handled non-parametrically; we refrain from specifying a model for it and do not need to estimate it.

Traditional work on DR typically assumes convergence of the estimates towards some 'least-false' parameter, both in the classical low dimension case and more recently also in high dimensions Farrell (2015); Tan (2018). We develop an approach based on the magnitude of the estimates instead without relying on the Lasso estimator to concentrate around some deterministic limit; dense high-dimensional models do not allow for existence of 'lease-false' direction. Our framework provides both a weaker requirement in theory and likely more applicable in practice.

In the following we begin with a special case of our estimator from Section 2, which has a closed-form expression and can be seen as directly estimating the hazards difference (HDi). In Section 3.2 we introduce the cross-fitting scheme, which leads to relaxation on sparsity assumptions. Finally we develop the doubly robust estimators in Section 3.3.

3.1 Hazards Difference (HDi) estimator

In Section 2 we allow for different estimators of the cumulative baseline hazard Λ under Theorem 1. Here we introduce a hazards difference (HDi) estimator that utilizes, a practically intuitive, covariate balancing.

Define a set of covariate balancing weights:

$$w_i^0(\gamma) = (1 - D_i)P(D_i = 1|Z_i), \;\; w_i^1(\gamma) = D_i P(D_i = 0|Z_i).$$

Remark 4. We say that the above weights are covariate balancing in the following sense. Consider the weighted empirical cumulative distribution functions of the covariates:

$$F_{d,n}(z)\frac{n^{-1}\sum_{i=1}^n w_i^d(\gamma_0)I(Z_i \le z)}{n^{-1}\sum_{i=1}^n w_i^d(\gamma_0)}, \; d = 0, 1,$$

where $I(Z \le z) = \prod_{j=1}^p I(Z^j \le z^j)$ with Z^j and z^j being the j-th component of Z and z, respectively. It is straightforward to show that $F_{d,n}(z)$ converges in probability to $\mathbb{E}\{\mathrm{Var}(D|Z)I(Z \le z)\}/\mathbb{E}\{\mathrm{Var}(D|Z)\}$ for $d = 0, 1$; i.e. the distributions of covariates in both treatment arms are approximately the same after weighting.

The role of covariate balancing has been discussed extensively in the recent literature (Hainmueller, 2012; Imai and Ratkovic, 2014; Zhao, 2019). The advantage of covariate balancing, in comparison to the traditional propensity score based approaches, is that it eliminates confounding even when the treatment model is misspecified, or when γ_0 cannot be estimated consistently (Athey et al., 2018; Bradic et al., 2019; Zubizarreta, 2015).

Define

$$\check{\Lambda}^k(t; \beta, \gamma) = \int_0^t \frac{\sum_{i=1}^n w_i^1(\gamma)\{dN_i(u) - Y_i(u)\beta^\top Z_i du\}}{\sum_{i=1}^n w_i^k(\gamma)Y_i(u)}, \qquad k = 0, 1. \tag{15}$$

Under the additive hazards model (1), $\check{\Lambda}^1$ and $\check{\Lambda}^0$ can be seen to estimate $\Lambda_0(t)+\theta t$ and $\Lambda_0(t)$, respectively. It is then immediate that the HDi estimator defined as

$$\check{\theta} = \frac{\sum_{i=1}^n \int_0^\tau w_i^0(\hat{\gamma})Y_i(t)d\left\{\check{\Lambda}^1(t; \hat{\beta}, \hat{\gamma}) - \check{\Lambda}^0(t; \hat{\beta}, \hat{\gamma})\right\}}{\sum_{i=1}^n w_i^0(\hat{\gamma})X_i}, \tag{16}$$

estimates θ under the additive hazards model (1).

In the following we show that under the logistic treatment model, $\check{\theta}$ is a special case of our estimators from Section 2, corresponding to a weighted Breslow estimator for Λ. In particular, define the weighted Breslow estimator as

$$\check{\Lambda}^k(t,\theta;\beta,\gamma) = \int_0^t \frac{\sum_{i=1}^n w_i^k(\gamma)\{dN_i(u) - Y_i(u)(D_i\theta + \beta^\top Z_i)du\}}{\sum_{i=1}^n w_i^k(\gamma)Y_i(u)}, \qquad k = 0, 1. \tag{17}$$

Note that the above weighted Breslow estimator satisfies Assumption 1 following a similar expansion like (14). As we show, the asymptotic variance of $\hat{\theta}$ does not depend on the specific estimator of the cumulative baseline hazard, so there is no loss of asymptotic efficiency by only using the treated subjects in $\check{\Lambda}$.

Using (17), the orthogonal score (6) becomes a linear function of θ:

$$\begin{aligned}
&\phi\big(\theta;\beta,\check{\Lambda}(\cdot,\theta;\beta,\gamma),\gamma\big)\\
=\;& -\frac{1}{n}\sum_{i=1}^n (1-D_i)\text{expit}(\gamma^\top\tilde{Z}_i)\int_0^\tau \Big(dN_i(u) - Y_i(u)\Big(\beta^\top\{Z_i - \check{Z}(u;\gamma)\}du + d\widetilde{N}(u;\gamma)\Big)\Big)\\
& -\frac{\theta}{n}\sum_{i=1}^n (1-D_i)\text{expit}(\gamma^\top\tilde{Z}_i)X_i,
\end{aligned} \tag{18}$$

where we used the weighted processes

$$\check{Z}(t;\gamma) = \sum_{i=1}^n Z_i w_i^1(\gamma)Y_i(t)/\sum_{i=1}^n w_i^1(\gamma)Y_i(t), \quad \text{and} \quad d\widetilde{N}(t;\gamma) = \sum_{i=1}^n w_i^1(\gamma)dN_i(t)/\sum_{i=1}^n w_i^1(\gamma)Y_i(t).$$

We then have

$$\check{\theta} = \frac{\sum_{i=1}^n w_i^0(\hat{\gamma})\int_0^\tau \Big(Y_i(u)\Big(\hat{\beta}^\top\{Z_i - \check{Z}(u;\hat{\gamma})\}du + d\widetilde{N}(u;\hat{\gamma})\Big) - dN_i(u)\Big)}{\sum_{i=1}^n w_i^0(\hat{\gamma})X_i}, \tag{19}$$

which is readily verified to be equal to (16).

When the additive hazards model (1) and the logistic regression model (3) are both correct, we may draw inference on θ using $\check{\theta}$ according to Theorem 1.

Theorem 2. *If Assumption 1-i - 1-vi hold and the rate condition*

$$\sqrt{\log(p)}\|\hat{\beta} - \beta_0\|_1 + \|\hat{\gamma} - \gamma_0\|_1 + \sqrt{n}\mathcal{D}_\gamma(\hat{\gamma},\gamma_0)\,\mathcal{D}_\beta\left(\hat{\beta},\beta_0\right) = o_p(1),$$

the HDi estimator $\check{\theta}$ satisfies

$$\hat{\sigma}^{-1}\sqrt{n}(\check{\theta} - \theta_0) \rightsquigarrow N(0,1),$$

where the variance estimator takes the closed form

$$\hat{\sigma}^2 = \frac{n^{-1}\sum_{i=1}^n \delta_i\{D_i - \text{expit}(\hat{\gamma}^\top\tilde{Z}_i)\}^2 e^{2\check{\theta}D_iX_i}}{\left\{n^{-1}\sum_{i=1}^n (1-D_i)\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)X_i\right\}^2}.$$

Note that double robustness in additive hazards models is complicated by the existence of a non-parametric nuisance parameter, $\Lambda(t)$. Leveraging the shape of the propensity surface while fitting the outcome model has generated considerable interest in recent years. We now propose to leverage the shape of the propensity surface while building an estimator of the additional non-parametric component as well. The key component is that now both the outcome model as well as the hazard leverage the shape of the propensity therefore, as we will show, achieving double robustness with weaker conditions.

3.2 Cross-fitting

Data-splitting (Cox, 1975) has been quite popular for the purpose of high-dimensional inference; see Athey et al. (2019); Chernozhukov et al. (2018a) and many more. Cross-fitting, is relatively new adaptation of data-splitting: there we estimate the nuisance parameters using all but the j-th fold of the data and evaluate the score on the j-fold of the data only. We then create a composite estimate by computing the averaged score across k-different data folds and solving for its root. In this way we hope to allow relaxed conditions on p, n and sparsity Algorithm 1 in the box demonstrates our cross-fitted orthogonal score method.

Data: split the data into k folds of equal size with the indices set $I_1, I_2, \ldots, I_k$
for *each fold indexed by* j **do**

1. estimate the nuisance parameters $\left(\hat{\beta}^{(j)}, \hat{\Lambda}^{(j)}, \hat{\gamma}^{(j)}\right)$ using the out-of-fold (out of j-th fold) data indexed by $I_{-j} = \{1, \ldots, n\} \setminus I_j$;
2. construct the cross-fitted score using the in-fold samples:

$$\phi^{(j)}\left(\theta; \hat{\beta}^{(j)}, \hat{\Lambda}^{(j)}, \hat{\gamma}^{(j)}\right) = \frac{1}{|I_j|} \sum_{i \in I_j} \left[D_i - \text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i)\right] \times \int_0^\tau e^{D_i \theta t} \left[dN_i(t) - Y_i(t) \left\{\left(D_i\theta + \hat{\beta}^{(j)\top} Z_i\right) dt + d\hat{\Lambda}^{(j)}(t;\theta)\right\}\right]. \quad (20)$$

end
Result: Obtain the estimated treatment effect $\hat{\theta}_{cf}$ by solving

$$\frac{1}{k} \sum_{j=1}^{k} \phi^{(j)}\left(\theta; \hat{\beta}^{(j)}, \hat{\Lambda}^{(j)}, \hat{\gamma}^{(j)}\right) = 0. \quad (21)$$

Algorithm 1: Composite estimation of the Treatment Effect via k-fold Cross-fitting

The cross-fitting algorithm described in Algorithm 1 induces independence between the score and the estimated nuisance parameters, further reducing the effect of the nuisance parameters on the treatment effect estimation; this is in addition to the orthogonality of the score function. For our purposes, the cross-fitting approach is also advantageous as it allows us to show theoretical gurantees that simple cross-validation suffices for selection of the tuning parameters; see Section 3.3.

Next, we introduce some additional notation. Let $(X_*, \delta_*, D_*, Z_*)$ be an independent copy as the original data, for which the expectation $\mathbb{E}_*$ is considered. We define the *average testing deviance* for the estimated model coefficients in (1) and (3):

$$\mathcal{D}^2_{\beta*}(\hat{\beta}, \beta_0) = \mathbb{E}_* \left[\int_0^\tau \left\{(\hat{\beta} - \beta_0)^\top Z_*\right\}^2 Y_*(t) dt\right], \ \mathcal{D}^2_{\gamma*}(\hat{\gamma}, \gamma_0) = \mathbb{E}_* \left[\left\{\text{expit}(\hat{\gamma}^\top Z_*) - \text{expit}(\gamma_0^\top Z_*)\right\}^2\right]. \quad (22)$$

Compared to the uniform error in Section 2, the average testing deviance has a convergence rate that grows slower when sparsity increases because of their connection with the estimation errors in l_2-norms,

$$\mathcal{D}_{\beta*}(\hat{\beta}, \beta_0) \leq \|\hat{\beta} - \beta_0\|_2 \sqrt{\sigma_{\max}\left\{\mathbb{E}_*(Z_* Z_*^\top)\right\} \tau} \leq \|\hat{\beta} - \beta_0\|_2 \sqrt{\left[\sigma_{\max}\left\{\text{Var}_*(Z_*)\right\} + K_Z^2\right] \tau},$$

and similarly $\mathcal{D}_{\gamma*}(\hat{\gamma}, \gamma_0) \leq \|\hat{\gamma} - \gamma_0\|_2 \sqrt{\sigma_{\max}\{\mathrm{Var}_*(Z_*)\} + K_Z^2}$, where $\sigma_{\max}(H)$ denotes the maximal eigenvalue of the matrix H.

Now we state our relaxed conditions, under which we have the inference result for θ using the cross-fitted estimator $\hat{\theta}_{cf}$ defined in (21), when both the additive hazards model and logistic regression model are correct.

Assumption 2. Suppose that Conditions 1-i to 1-viii hold with $(\hat{\beta}, \hat{\Lambda}, \hat{\gamma}) = \left(\hat{\beta}^{(j)}, \hat{\Lambda}^{(j)}, \hat{\gamma}^{(j)}\right)$ for all $j = 1, \ldots, k$. Assume additionally,

(i) the rates of estimation errors follow

$$\mathcal{D}_{\beta*}\left(\hat{\beta}^{(j)}, \beta_0\right) + \sup_{t\in[0,\tau]}\left|\hat{\Lambda}^{(j)}(t;\theta_0) - \Lambda_0(t)\right| + \mathcal{D}_{\gamma*}\left(\hat{\gamma}^{(j)}, \gamma_0\right)$$
$$+\sqrt{n}\mathcal{D}_{\gamma*}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\left(\mathcal{D}_{\beta*}\left(\hat{\beta}^{(j)}, \beta_0\right) + \sup_{t\in[0,\tau]}\left|\hat{\Lambda}^{(j)}(t;\theta_0) - \Lambda_0(t)\right|\right) = o_p(1). \tag{23}$$

Theorem 3. *Under Assumption 2, $\hat{\theta}_{cf}$ obtained from Algorithm 1 satisfies*

$$\hat{\sigma}_{cf}^{-1}\sqrt{n}(\hat{\theta}_{cf} - \theta_0) \rightsquigarrow N(0,1),$$

with the closed-form variance estimator

$$\hat{\sigma}_{cf}^2 = \frac{n^{-1}\sum_{j=1}^k \sum_{i\in I_j} \delta_i\{D_i - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}^2 e^{2\hat{\theta}D_iX_i}}{\left\{n^{-1}\sum_{j=1}^k \sum_{i\in I_j}(1-D_i)\mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)X_i\right\}^2}.$$

We make a few remarks for the above result. The cross-fitted score (21) can handle a larger number of covariates, less sparse models and less restrictive estimators of the baseline hazard without Assumptions 1-vii and 1-viii. The removal of condition (10) allows various estimation methods of the baseline hazards besides the Breslow type estimators, e.g. splines. In addition, (23) allows a larger dimension without the extra $\log(p)$ factor in (9). Lastly, interpreting (23) we observe that a sufficient condition for Theorem 3 is

$$s_\beta = o(n/\log(p)),\ s_\gamma = o(n/\log(p)),\ \text{and}\ s_\beta s_\gamma = o\left(n/\log(p)^2\right),$$

which is slightly weaker than those discussed in Section 2. This product rate condition is the same as that of Chernozhukov et al. (2018a) derived for fully observed data without censoring. We hence remark that our obtained results achieve the certain optimality in rate condition, considering the fact that our model has an additional nonparametric component arising from the presence of the baseline hazard.

3.3 Model and Sparsity double robustness

In this subsection we first consider *model double robustness*, where either the outcome or the treatment model can be misspecified. We then consider *sparsity double robustness*, where the sparsity assumption for either model might be violated.

Here we consider and the general treatment model (7) and the partially linear additive hazards model (8). Note that θ as defined in (8) retains the treatment effect interpretation.

3.3.1 Magnitude of cross-validation

We define the *magnitude of estimation* under the possibly misspecified model

$$\mathcal{M}_\beta^2(\hat{\beta}) = \int_0^\tau \hat{\beta}^\top \mathbb{E}_* \left[\{Z_* - \mu(t)\}^{\otimes 2} Y_*(t) \right] \hat{\beta} dt,$$
$$\mathcal{M}_\gamma(\hat{\gamma}) = \left[\mathbb{E}_* \left\{ w_*^0(\hat{\gamma}) X_* \right\} \right]^{-1} + \left[\mathbb{E}_* \left\{ w_*^1(\hat{\gamma}) Y_*(\tau) \right\} \right]^{-1} \tag{24}$$

with $\mu(t) = \mathbb{E}_*(Z_*)/\mathbb{E}_*\{Y_*(t)\}$. Notion of magnitude can be understood through

$$\mathcal{M}_\beta(\hat{\beta}) \le \|\hat{\beta}\|_1 \|Z_*\|_\infty \sqrt{\tau}, \quad \mathcal{M}_\gamma(\hat{\gamma}) \le \frac{1 + e^{\|\hat{\gamma}\|_1 K_Z}}{\tau \mathbb{E}_*\{(1 - D_*) Y_*(\tau)\}} + \frac{1 + e^{\|\hat{\gamma}\|_1 K_Z}}{\mathbb{E}_*\{D_* Y_*(\tau)\}};$$

with a finite $\|Z_*\|_\infty$ and strictly positive $\mathbb{E}_*\{D_* Y_*(\tau)\}$ and $\mathbb{E}_*\{(1 - D_*) Y_*(\tau)\}$, $\mathcal{M}_\beta(\hat{\beta})$ and $\mathcal{M}_\gamma(\hat{\gamma})$ are guaranteed to be finite as long as both l^1-norms are finite.

We will make an assumption that the magnitudes above are bounded. Below we show that using cross-validation to select the regularization tuning parameters is sufficient to control the introduced magnitudes. Let $\{\hat{\beta}(\lambda) : \lambda > 0\}$ and $\{\hat{\gamma}(\lambda) : \lambda > 0\}$ be classes of Lasso estimators with different penalty factors λ under additive hazards model and logistic regression model, respectively. The sets are often called the Lasso regularization path (Friedman et al., 2010). We consider risk minimization to choose penalty factors corresponding to cross-validation:

$$\hat{\lambda}_\beta = \underset{\lambda > 0}{\operatorname{argmin}}\, l_\beta^*(\hat{\beta}(\lambda)), \;\; \hat{\lambda}_\gamma = \underset{\lambda > 0}{\operatorname{argmin}}\, l_\gamma^*(\hat{\gamma}(\lambda)), \tag{25}$$

where the generalization losses for β and γ are defined as

$$l_\beta^*(\beta) = \int_0^\tau \mathbb{E}_* \left(\left[\beta^\top \{Z_* - \mu(t)\} \right]^2 Y_*(t) \right) dt - 2 \int_0^\tau \mathbb{E}_* \left[\beta^\top \{Z_* - \mu(t)\} \, dN_*(t) \right],$$
$$l_\gamma^*(\gamma) = -\mathbb{E}_*(D_* \gamma^\top Z_*) + \mathbb{E}_* \left\{ \log \left(1 + e^{\gamma^\top Z_*} \right) \right\}. \tag{26}$$

Lemma 4. *Under the partially linear additive hazards model (8), we have*

$$\mathcal{M}_\beta^2 \left(\hat{\beta}(\hat{\lambda}_\beta) \right) \le 4 \int_0^\tau \mathbb{E}_*\{g(t, Z_*)^2 Y_*(t)\} dt.$$

Lemma 5. *Under Assumption 1-v where ε_Y is defined, we have with probability tending to one*

$$\mathcal{M}_\gamma \left(\hat{\gamma}(\hat{\lambda}_\gamma) \right) \le (1 + \tau^{-1}) 8 \varepsilon_Y^{-3} e^{-4 \log(\varepsilon_Y)/\varepsilon_Y^2}.$$

Lemmas 4 and 5 give surprisingly simple guarantees on the cross-validated Lasso estimators even when the model assumptions are wrong. Our results here have opened up a possibly new direction of utilizing penalization methods under model misspecification. Unlike the traditional "least false parameter" argument (Farrell, 2015; Hjort, 1986, 1992; Tan, 2018), our bounds on the magnitude require no specific assumptions on how the model is misspecified, including sparsity, restricted eigenvalues, etc.. Consequently, our doubly robust estimation developed under the magnitude conditions is stronger in theory than existing results and conditions more likely satisfied in applications.

3.3.2 Model double robustness

Here we apply the cross-fitting algorithm described in Section 3.2 to the score (18) introduced in Section 3.1, i.e., to the HDi estimator. Suppose that $(\hat{\beta}^{(j)}, \hat{\gamma}^{(j)})$ is the Lasso estimator of (β, γ) using the out-of-fold data for the fold j. The cross-fitted version of (19) takes the following form:

$$\check{\theta}_{cf} = \sum_{j=1}^{k} \sum_{i \in I_j} \frac{\int_0^\tau w_i^0 \left(\hat{\gamma}^{(j)}\right) Y_i(t)}{\sum_{j=1}^{k} \sum_{i \in I_j} w_i^0 \left(\hat{\gamma}^{(j)}\right) X_i} d \left\{ \check{\Lambda}^{1(j)} \left(t; \hat{\beta}^{(j)}, \hat{\gamma}^{(j)}\right) - \check{\Lambda}^{0(j)} \left(t; \hat{\beta}^{(j)}, \hat{\gamma}^{(j)}\right) \right\}, \tag{27}$$

with $\check{\Lambda}^{k(j)}(t; \beta, \gamma)$, for $k = 0, 1$, defined similarly as in (15) but now on a j-th fold only

$$\check{\Lambda}^{k(j)}(t; \beta, \gamma) = \int_0^t \frac{\sum_{i \in I_j} w_i^k(\gamma) \{dN_i(u) - Y_i(u) \beta^\top Z_i du\}}{\sum_{i \in I_j} w_i^1(\gamma) Y_i(u)}, \qquad k = 0, 1.$$

Assumption 3.

(a)- Suppose that Conditions 1-ii to 1-iv in Assumption 1 are satisfied, as well as

(i) the outcome follows additive hazards (1) whereas the treatment follows the general model (7);

(ii) $\sup_{j=1,\ldots,k} \mathcal{D}_{\beta *} \left(\hat{\beta}^{(j)}, \beta_0\right) = o_p(1)$ and $\sup_{j=1,\ldots,k} \mathcal{M}_\gamma \left(\hat{\gamma}^{(j)}\right) \leq K_{\mathcal{M}_g} + o_p(1)$.

Or:

(b)- Suppose that Conditions 1-ii, 1-iii and 1-v in Assumption 1 are satisfied, as well as

(i) the outcome follows partially linear additive hazards (8) whereas the treatment follows the logistic regression model (3);

(ii) $g(t; Z)$ satisfies $\mathbb{E}\left\{\int_0^\tau g^2(t; Z_i) Y_i(t) dt\right\} = K_\Lambda^2 = o(n)$;

(iii) the rate condition holds

$$\left\{ K_\Lambda + \sup_{j=1,\ldots,k} \mathcal{M}_\beta \left(\hat{\beta}^{(j)}\right) \right\} \sup_{j=1,\ldots,k} \mathcal{D}_{\gamma *} \left(\hat{\gamma}^{(j)}, \gamma_0\right) = o_p(1).$$

Theorem 4. *(Model double robustness) When either Assumption 3(a) or 3(b) holds, $\check{\theta}_{cf}$, defined in (27), is consistent for θ_0, i.e. $|\check{\theta} - \theta_0| = o_p(1)$ when $p, n \to \infty$.*

Assumption 3 describes the regularity conditions for doubly robust estimation. Under Assumption 3(a), the additive hazards model is correct for β_0 while the propensity model is misspecified. Besides the regularity Conditions 1-ii to 1-iv in Assumption 1, we only ask for a finite magnitude of the estimator of the propensity model $\hat{\gamma}$. According to Lemma 5, such bound is guaranteed by cross-validation if we assume additionally the Condition 1-v in Assumption 1.

Under Assumption 3(b), the propensity model is correct with γ_0, while the outcome model, i.e, the additive hazards model is misspecified. Besides the regularity Conditions 1-ii, 1-iii and 1-v in Assumption 1, we actually allow the average contribution of covariate in hazard K_Λ and the magnitude of the estimator of the additive hazards model $\hat{\beta}$ to grow with sample size, as long as their rate of growth is smaller than the rate of convergence of $\hat{\gamma}$. In fact, it is quite common (Hou et al., 2019; Yu et al., 2019) in high dimensions to assume a uniform bound on the hazard, i.e. $\sup_{t \in [0,\tau]} \sup_{i=1,\ldots,n} g(t, Z_i)$, which is much stronger than

our condition. According to Lemma 4, the magnitude of $\hat{\beta}$ must grow slower than K_Λ when the penalty for $\hat{\beta}$ is selected by cross-validation.

Our proof relies on the double robustness of our score at population level. When the additive hazards model (1) is correct, the true θ_0 solves the equation $\mathbb{E}\left\{\phi^{(j)}\left(\theta_0; \beta_0, \Lambda_0, \hat{\gamma}^{(j)}\right)\right\} = 0$ for any $\hat{\gamma}^{(j)}$. When the logistic regression model (3) is correct, the true θ_0 solves the equation $\mathbb{E}\left\{\phi^{(j)}\left(\theta_0; \hat{\beta}^{(j)}, \hat{\Lambda}^{(j)}, \gamma_0\right)\right\} = 0$ with any $\hat{\beta}^{(j)}$ and $\hat{\Lambda}^{(j)}$.

3.3.3 Sparsity double robustness

We now turn our focus to sparsity double robustness. Now either the outcome or the propensity (but not necessarily both) are defined through sparse (and not necessarily ultra-sparse) high-dimensional models. Here, ultra sparse rate denotes that of $o(\sqrt{n/\log(p)})$ where sparse denotes that of $o(n/\log(p))$. This setup is far more general that those of previous sections as either outcome or propensity model can be defined through a fully dense high-dimensional model.

Under Assumption 1-i, both the additive hazards model and the logistic regression model are correct. Different from Theorem 1 in which both true coefficients are assumed sparse, Theorem 5 only requires one of them to be sparse; the other one can be as large as p and possibly much larger than n, as we only require

$$\min\{s_\beta, s_\gamma\} = o(n/\log(p)),$$

a condition that is extremely generous in high-dimensional setting.

Theorem 5. *(Sparsity double robustness) Suppose we use Lasso (11) and (12) to estimate $(\hat{\beta}^{(j)}, \hat{\gamma}^{(j)})$ in k-fold cross-fitting. The penalty factors are selected by minimizing the generalization loss (25). Under the Conditions 1-i to 1-v in Assumption 1 and additionally:*

(i) the dimensions satisfy $\min\{s_\beta, s_\gamma\} = o(n/\log(p))$,

(ii) for the model with the smaller sparsity, the minimal eigenvalue of the population Hessian is bounded away from zero,

$$\sigma_{\min}\left(\mathbb{E}\left[\int_0^\tau \begin{pmatrix} D - \mathbb{E}\{DY(t)\}/\mathbb{E}\{Y(t)\} \\ Z - \mathbb{E}\{ZY(t)\}/\mathbb{E}\{Y(t)\} \end{pmatrix}^{\otimes 2} Y(t)dt\right]\right) > \varepsilon_\Sigma, \qquad \text{if } s_\beta \le s_\gamma;$$

$$\sigma_{\min}\left(\mathbb{E}\left\{\mathrm{Var}(D|Z)Z^{\otimes 2}\right\}\right) > \varepsilon_\Sigma, \qquad \text{if } s_\beta > s_\gamma,$$

(iii) $\sigma_{\max}\{\mathrm{Var}(Z)\} < K_Z^2$,

(iv) $\int_0^\tau \mathbb{E}_*\{(\beta_0^\top Z_*)^2 Y_*(t)\}dt < K_\Lambda \asymp 1$,

then, $\check{\theta}_{cf}$, defined in (27), is consistent for θ_0, i.e. $|\check{\theta}_{cf} - \theta_0| = o_p(1)$, when $p, n \to \infty$.

Remark 5. Theorem 5 demonstrates our unique contribution to the doubly robust estimation in high dimensions. When the sparsity exceeds the sample size, various concentration results on the penalized estimators no longer hold. Consequently, doubly robust estimation results established with the limit of nuisance parameters (Farrell, 2015; Tan, 2018) do not apply. Regardless, we are able to show that our orthogonal score method produces consistent estimation, while utilizing simple cross-validation.

We note that inference on θ using $\check{\theta}_{cf}$ is fundamentally different from the existing work that considered only low-dimensional covariates (Jiang et al., 2017; Kang et al., 2018; Wang et al., 2017; Zhang and Schaubel, 2012; Zhao et al., 2015), and is beyond the scope of the current paper.

4 Simulation

We assessed the performance of the proposed estimators in the following simulation. We considered $n = p = 300$ and $n = p = 1500$. To ensure that the baseline hazard is non-negative, the covariates $Z_1, \ldots, Z_p$ were independently generated from $N(0,1)$ conditioned on the event that $\beta^\top Z_i \geq 0.25$. The censoring time C was generated as the smaller between τ and Uniform $(0, c_0)$; the parameters τ and c_0 were chosen such that $n/10$ treated subjects were expected to be at-risk at $t = \tau$, and the censoring rate was around 30%. We repeated the simulation 500 times.

We considered the four proposed estimators: $\hat{\theta}$ obtained from (6) with the Breslow estimator (13), the HDi estimator $\check{\theta}$ in (19), and their cross-fitted counterparts $\hat{\theta}_{cf}$ as described in Algorithm 1 and $\check{\theta}_{cf}$ in (27). For comparison, we also present the result of the Lasso estimator $\tilde{\theta}$ under the additive hazards model with the penalty for θ set to be zero; its estimation bias is then entirely caused by regularization on the nuisance parameters.

We estimated the coeffcient of the logistic regression with covariates Z by the Lasso estimator $\hat{\gamma}$ in (12) using the R-package *glmnet*, and the coefficient of the additive hazards model with covariates (D, Z) by the Lasso estimator $(\hat{\theta}_l, \hat{\beta})$ in (11) using the R-package *ahaz*. For $\hat{\theta}$ and $\check{\theta}$ the penalty parameters were selected by 10-fold cross-validation. For cross-fitting the number of folds was set as 10, and within each fold we used 9-fold cross-validation for Lasso.

4.1 Inference under correctly specified models

We first generated the event time T from the additive hazards model (1) with $\theta_0 = -0.25$, and $\lambda_0(t) = 0.25$. We generated the treatment assignment D from the logistic regression model (3); the intercept in the treatment model was always chosen so that the marginal probability $\mathbb{P}(D = 1) = 0.5$. We considered the following three sparsity levels:

$$\text{very sparse } s_\beta = 2\text{: } \beta = (1, 0.1, \underbrace{0, \ldots, 0}_{p-2}), \quad s_\gamma = 1 : \gamma = (1, \underbrace{0, \ldots, 0}_{p-1});$$

$$\text{sparse } s_\beta = 6\text{: } \beta = (1, \underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0, \ldots, 0}_{p-6}), \quad s_\gamma = 3 : \gamma = (1, 0.05, 0.05, \underbrace{0, \ldots, 0}_{p-3});$$

$$\text{moderately sparse } s_\beta = 15\text{: } \beta = (1, \underbrace{0.1, \ldots, 0.1}_{13}, \underbrace{0, \ldots, 0}_{p-15}),$$

$$s_\gamma = 10 : \gamma = (1, 1, \underbrace{0.05, \ldots, 0.05}_{8}, \underbrace{0, \ldots, 0}_{p-10}).$$

We set the sparsities of the logistic regression model to be smaller than those of the additive hazards model because the Lasso estimator under the former is empirically more sensitive to increase in sparsity. Four pairs of sparsities, $(s_\beta = 2, s_\gamma = 1)$, $(s_\beta = 2, s_\gamma = 10)$, $(s_\beta = 15, s_\gamma = 1)$ and $(s_\beta = 6, s_\gamma = 3)$, were studied in the simulation. In the above we considered two types of signals: a strong signal of size 1 for both β and γ, and a weak signal of sizes $\beta = 0.1$ and $\gamma = 0.05$, respectively.

We present the inference results in Table 1. The naive $\tilde{\theta}$ had large bias/SD ratio, which confirmed the difficulty in drawing inference. In comparison, the orthogonal score approach largely reduced the bias in all four proposed estimators, especially for the larger sample size $n = p = 1500$. All four estimators achieved reasonably good coverage rates of the nominal 95% confidence intervals with the larger sample size $n = p = 1500$ with cross-fitted HDi estimator outperforming in small sample sizes. With the smaller sample size $n = p = 300$, $\check{\theta}_{cf}$ the cross-fitted HDi estimator also had good coverage properties; this appears

Table 1: Inference results under correctly specified models: true $\theta = -0.25$, 30% censoring. $\hat{\theta}$, $\check{\theta}$, $\hat{\theta}_{cf}$ and $\check{\theta}_{cf}$ are the four proposed estimators, where the subscript 'cf' denotes the cross-fitted version. The naive Lasso estimator $\tilde{\theta}$ penalized only the covariate effects β but not θ. 'CP' is the coverage probability of the nominal 95% confidence intervals.

Sparsity		Lasso $\tilde{\theta}$		$\hat{\theta}$				HDi $\check{\theta}$			
s_β	s_γ	Bias	SD	Bias	SD	SE	CP	Bias	SD	SE	CP
n=p=300											
2	1	0.054	0.097	0.029	0.097	0.091	92.4 %	0.032	0.094	0.091	93.0 %
6	3	0.071	0.094	0.050	0.095	0.093	92.0 %	0.052	0.092	0.093	93.0 %
15	1	0.088	0.135	0.051	0.123	0.128	93.6 %	0.049	0.122	0.128	94.0 %
2	10	0.099	0.094	0.050	0.099	0.094	89.6 %	0.052	0.096	0.094	89.8 %
n=p=1500											
2	1	0.031	0.040	0.009	0.041	0.041	94.2 %	0.011	0.041	0.041	94.0 %
6	3	0.033	0.042	0.015	0.043	0.042	93.0 %	0.017	0.042	0.042	93.6 %
15	1	0.047	0.063	0.019	0.064	0.058	91.4 %	0.020	0.063	0.058	91.2 %
2	10	0.077	0.041	0.019	0.043	0.043	91.6 %	0.022	0.043	0.043	91.6 %

Sparsity		$\hat{\theta}_{cf}$				HDi $\check{\theta}_{cf}$			
s_β	s_γ	Bias	SD	SE	CP	Bias	SD	SE	CP
n=p=300									
2	1	0.012	0.100	0.090	91.0 %	0.011	0.093	0.090	93.4 %
6	3	0.027	0.100	0.092	92.4 %	0.028	0.089	0.092	94.6 %
15	1	0.018	0.134	0.127	93.8 %	0.013	0.123	0.127	95.8 %
2	10	0.032	0.106	0.094	89.4 %	0.032	0.097	0.094	93.2 %
n=p=1500									
2	1	0.006	0.042	0.041	94.8 %	0.009	0.040	0.041	95.4 %
6	3	0.010	0.044	0.041	92.8 %	0.014	0.041	0.042	94.2 %
15	1	0.006	0.064	0.058	92.4 %	0.012	0.061	0.058	93.0 %
2	10	0.017	0.044	0.043	92.4 %	0.019	0.042	0.043	92.8 %

due to both less bias than the HDi estimator $\check{\theta}$, as well as closer approximation of SE to SD compared with $\hat{\theta}_{cf}$.

4.2 Double robustness

We first considered sparsity DR. We simulated from the models above with dense coefficients:

$$\text{Dense } s_\beta = 30\text{: } \beta = (\underbrace{1,\ldots,1}_{4},\underbrace{0.1,\ldots,0.1}_{26},\underbrace{0,\ldots,0}_{p-30}),$$
$$\text{Dense } s_\gamma = 20\text{: } \gamma = (\underbrace{1,\ldots,1}_{4},\underbrace{0.05,\ldots,0.05}_{16},\underbrace{0,\ldots,0}_{p-20}).$$

Two pairs of very sparse - dense combinations, $(s_\beta = 2, s_\gamma = 20)$ and $(s_\beta = 30, s_\gamma = 1)$, were studied.

We then considered model DR in three different scenarios, denoted by 'E', 'P' and 'D' which stand for 'Exponential', 'Probit' and 'Deterministic', respectively.

- In scenario 'E', the event time was generated with exponential link:

$$\lambda(t|D,Z) = -0.25D + \exp\left(\beta^\top Z\right) + 0.25, \tag{28}$$

  while the logistic treatment model (3) was correct. The coefficients were set as in $(s_\beta = 2, s_\gamma = 1)$.

- In scenario 'P', we considered the misspecified treatment model with probit link:

$$\mathbb{P}(D = 1|Z) = \text{probit}\left(\gamma^\top \tilde{\text{Z}}\right), \tag{29}$$

  while the additive hazards model (1) for the event time was correct. The coefficients were also set as in $(s_\beta = 2, s_\gamma = 1)$.

- In scenario 'D', we considered another misspecified treatment model with deterministic treatment assignment given Z:

$$D|Z = I\left(\gamma^\top \tilde{Z} > \mu\right), \tag{30}$$

  where μ is the median of $\gamma^\top \tilde{Z}$. Again the additive hazards model (1) for the event time was correct, and the coefficients were set as in $(s_\beta = 2, s_\gamma = 1)$. Since $\text{Var}(D|Z) = 0$ for all Z, i.e. there was no overlap at all, Assumptions 1-v and 1-vi were violated in this scenario.

We present the estimation results in Table 2. All four proposed estimators had reasonable estimation errors that decayed with increased sample size, showing evidence of consistency. All four estimators had smaller bias than the naive Lasso $\tilde{\theta}$, most notably under scenario 'D' where the naive Lasso failed completely. Like in Table 1, $\hat{\theta}_{cf}$ and $\check{\theta}_{cf}$ demonstrated the advantage of cross-fitting with even smaller biases. Under most scenarios, the HDi $\check{\theta}_{cf}$ had smaller SD than $\hat{\theta}_{cf}$ when $n = p = 300$, which leaded to the improved MSE. Again in the scenario 'D' with $n = p = 300$, $\check{\theta}_{cf}$ showed advantage over $\hat{\theta}_{cf}$ which could not be defined in one of the simulation runs where the score equation appeared to have no root; on the other hand, the closed-form HDi had no such issues.

To further understand the behavior of the proposed estimators as seen above, we empirically investigated the average testing deviance and the magnitude of estimation described in Section 3. We used the

Table 2: Doubly robust results under dense coefficients ($s_\beta = 30$, $s_\gamma = 20$) or misspecified models (exponential link 'E', probit link 'P' and deterministic treatment assignment 'D'). True $\theta = -0.25$, 30% censoring. $\hat{\theta}$, $\check{\theta}$, $\hat{\theta}_{cf}$ and $\check{\theta}_{cf}$ are the four proposed estimators, where the subscript 'cf' denotes the cross-fitted version. The naive Lasso estimator $\tilde{\theta}$ penalized only the covariate effects β but not θ.

Sparsity		Lasso $\tilde{\theta}$			$\hat{\theta}$			HDi $\check{\theta}$		
s_β	s_γ	Bias	sd	$\sqrt{MSE}$	Bias	sd	$\sqrt{MSE}$	Bias	sd	$\sqrt{MSE}$
n=p=300										
30	1	0.141	0.202	0.247	0.080	0.169	0.187	0.074	0.169	0.185
2	20	0.078	0.090	0.119	0.051	0.099	0.111	0.052	0.096	0.109
E	1	0.233	0.397	0.461	0.117	0.375	0.393	0.106	0.366	0.381
2	P	0.095	0.102	0.139	0.041	0.101	0.109	0.043	0.097	0.106
2	D	-0.598	0.204	0.632	-0.103	0.217	0.240	-0.124	0.223	0.255
n=p=1500										
30	1	0.057	0.083	0.101	0.018	0.082	0.084	0.018	0.081	0.083
2	20	0.061	0.042	0.074	0.020	0.048	0.052	0.022	0.047	0.052
E	1	0.132	0.169	0.214	0.049	0.167	0.174	0.049	0.163	0.171
2	P	0.050	0.043	0.066	0.012	0.045	0.046	0.014	0.044	0.046
2	D	-0.308	0.092	0.321	-0.032	0.104	0.109	-0.040	0.107	0.114

Sparsity		$\hat{\theta}_{cf}$			HDi $\check{\theta}_{cf}$		
s_β	s_γ	Bias	sd	$\sqrt{MSE}$	Bias	sd	$\sqrt{MSE}$
n=p=300							
30	1	0.049	0.197	0.203	0.026	0.177	0.179
2	20	0.036	0.106	0.112	0.034	0.099	0.104
E	1	0.054	0.396	0.400	0.001	0.364	0.364
2	P	0.021	0.105	0.107	0.019	0.097	0.099
2	D	-0.216*	0.506*	0.550*	-0.133	0.258	0.290
n=p=1500							
30	1	0.000	0.085	0.085	0.004	0.080	0.080
2	20	0.018	0.049	0.052	0.020	0.047	0.051
E	1	0.027	0.169	0.171	0.024	0.163	0.165
2	P	0.008	0.045	0.046	0.011	0.043	0.045
2	D	-0.031	0.107	0.111	-0.040	0.116	0.123

*One divergent iteration was removed from the summary.

sample average $n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}$ to approximate the expectation $\mathbb{E}_*$ for the former, so that

$$\widehat{\mathcal{D}_{\beta*}}^2 = n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}\left[\int_0^\tau \left\{(\hat{\beta}^{(j)}-\beta_0)^\top Z_i\right\}^2 Y_i(t)dt\right],$$
$$\widehat{\mathcal{D}_{\gamma*}}^2 = n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}\left[\left\{\text{expit}\left(\hat{\gamma}^{(j)\top}Z_i\right)-\mathbb{E}\left(D_i|Z_i\right)\right\}^2\right]. \tag{31}$$

Under the Scenario 'E', we evaluated the average testing deviance by comparing the predicted contributions of covariates in the score with the true contributions:

$$\widehat{\mathcal{D}_{\beta*}} = n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}\left[\int_0^\tau \left\{\hat{\beta}^{(j)\top}Z_i-\exp(\beta_0^\top Z_i)\right\}^2 Y_i(t)dt\right]. \tag{32}$$

For the latter we used the sample average $k/n\sum_{i\in I_k}$ to approximate the expectation $\mathbb{E}_*$, and took the maximum across all folds, so that

$$\widehat{\mathcal{M}_\beta} = \max_{j=1,\ldots,k}\sqrt{\int_0^\tau \hat{\beta}^{(j)\top}\frac{k}{n}\sum_{i\in I_j}\left[\left\{Z_i-\bar{Z}^{(j)}(t)\right\}^{\otimes 2}Y_i(t)\right]\hat{\beta}^{(j)}dt},$$
$$\widehat{\mathcal{M}_\gamma} = \max_{j=1,\ldots,k}\left\{\frac{n/k}{\sum_{i\in I_j}w_i^0\left(\hat{\beta}^{(j)}\right)X_i}+\frac{n/k}{\sum_{i\in I_j}w_i^1\left(\hat{\beta}^{(j)}\right)Y_i(\tau)}\right\}. \tag{33}$$

In Table 3, we present the estimation error, deviance and magnitude of the nuisance parameters. The uniform error columns contain the estimation errors from Lasso in l_1-norm and the Breslow estimator in l_∞-norm; these are compared to the true parameters when the models are correctly specified. The deviance columns contain the mean over simulation runs of the empirical deviance defined in (31) and (32). The magnitude columns contain the median over simulation runs of the empirical magnitudes defined in (33).

When the true model is dense, in our setups either $s_\beta = 30$ or $s_\gamma = 20$, we observe from Table 3 that the Lasso estimators deviated substantially from the underlying true coefficients in l^1-norm, suggesting that the Lasso was not concentrated around the true parameters. Regardless, our proposed method achieved consistent estimation of the treatment effect. When the Assumption 1-v held, i.e. for all scenarios except 'D', the magnitudes were well controlled, with no indication that the magnitudes might blow up in high dimensions.

5 Data Analysis

Clinical databases such as the United States National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) typically contain disease specific variables, but with only limited information on the subjects' health status such as comorbidities. In studying causal treatment effects, this can lead to unobserved confounding (Hadley et al., 2010; Ying et al., 2019). On the other hand, the availability of information from insurance claims databases could make up for some of these otherwise 'unobserved' variables, and they have been shown to contain much information about these comorbidities (Hou et al., 2018; Riviere et al., 2019).

For prostate cancer while radical prostatectomy is quite effective in reducing cancer related deaths, with improvement in diagnosis, treatment and management for the disease other causes have become

Table 3: Estimation error, deviance and magnitude of the nuisance parameters; the settings are as described for the previous two tables.

	Outcome Model					Treatment Model			
n & p	s_β	$\hat{\beta}$ in l_1	$\hat{\Lambda}$ in l_∞	$\widehat{\mathcal{D}_{\beta*}}(\check{\beta})$	$\widehat{\mathcal{M}_\beta}(\check{\beta})$	s_γ	$\hat{\gamma}$ in l_1	$\widehat{\mathcal{D}_{\gamma*}}(\check{\gamma})$	$\widehat{\mathcal{M}_\gamma}(\check{\gamma})$
300	2	0.61	0.27	0.34	0.37	1	1.38	0.09	36.73
1500	2	0.42	0.13	0.20	0.41	1	0.80	0.05	29.78
300	6	1.13	0.31	0.46	0.35	3	1.75	0.10	29.84
1500	6	0.90	0.17	0.27	0.40	3	1.13	0.05	23.49
300	15	2.87	0.46	0.67	0.38	10	2.97	0.12	35.11
1500	15	2.43	0.28	0.45	0.44	10	2.02	0.07	27.16
300	30	6.22	0.60	0.96	0.42	20	5.01	0.15	48.72
1500	30	5.19	0.36	0.63	0.48	20	3.75	0.09	34.97
300	E	–	–	0.71	0.58	P	–	0.09	29.19
1500	E	–	–	0.47	0.65	P	–	0.05	22.91
300						D	–	0.17	$> 100^{**}$
1500						D	–	0.13	$> 100^{**}$

* The dashed entries are not well-defined due to misspecification;
** The divergence of magnitude is expected because setup "D" violates Assumption 1-v.

dominant for the overall death of this patient population (Lu-Yao et al., 2004). Such changes suggest a comprehensive consideration for medical decisions on the initial treatment. Studying the comparative effect of radical prostatectomy on overall survival of the patient using observational data calls for proper control of confounding. Due to the lack of tools for handling the high-dimensional claims code data, existing work on the topic either have not made use (Satkunasivam et al., 2015; Ying et al., 2019), or made very limited through summary statistics (Hadley et al., 2010), of the rich information on the patients' health status.

Motivated by our previous linked SEER-Medicare database projects, we considered 49973 prostate cancer patients diagnosed during 2004-2013 as recorded in the SEER-Medicare linked database. The data contained the survival times of the patients, treatment information, demographic information, clinical variables and the federal Medicare insurance claims codes. More specifically, we included in our analysis age, race, martital status, tumor stage, tumor grade, prior Charlson comorbidity score, and 6397 claims codes possessed by at least 10 patients during the 12 months before their diagnosis of prostate cancer. The summary statistics of these variables are presented in Table 4. Our main focus is the treatment effect of surgery (radical prostatectomy) on the overall survival of the patients. In our sample, 17614 (35.25 %) patients received surgery while 32359 (64.75 %) patients received other types of treatment without surgery. As can be seen, many of the variables are not balanced between the surgery and no surgery groups; in particular, patients who received surgery tended to be younger, married, white, T2, poorly differentiated tumor grade, zero comorbidity, and diagnosed in 2008 or earlier. Among all patients 5375 (10.76 %) deaths were observed while the rest of the patients were still alive by the end of year 2013. The Kaplan-Meier curves for the two groups are presented in Figure 1.

With the methods proposed in this paper, we studied the treatment effect of radical prostatectomy on patient survival using the high-dimensional covariates. The causal diagrams of the analyses are illustrated in Figure 2. In Analysis I, we adjusted for the potential confounding effects from the clinical and demographic variables and the high-dimensional claims codes. After excluding claims codes with less than 10 occurrences, we had 6533 covariates in Analysis I. For the additive hazards model, we were under the '$p > n$' scenario as the number of covariates exceeded the number of observed events 5375.

We applied the same methodology as implemented in the simulation Section. In Analysis I, Lasso

Table 4: Summary of the linked SEER-Medicare data

		No Surgery	Surgery
Variable	Value	$n = 32359$	$n = 17614$
Age	66-69	12460 (38.5 %)	8790 (49.9 %)
	70-74	19899 (61.5 %)	8824 (50.1 %)
Marital status	Married	21464 (66.3 %)	13439 (76.3 %)
	Divorced	2200 (6.8 %)	820 (4.7 %)
	Single	2490 (7.7 %)	1207 (6.9 %)
	Other	6205 (19.2 %)	2148 (12.2 %)
Race	White	26019 (80.4 %)	15035 (85.4 %)
	Black	4501 (13.9 %)	1527 (8.7 %)
	Asian	467 (1.4 %)	284 (1.6 %)
	Hispanic	327 (1.0 %)	204 (1.2 %)
	Other	1045 (3.2 %)	564 (3.2 %)
Tumor stage	T1	20314 (62.8 %)	3866 (21.9 %)
	T2	12045 (37.2 %)	13748 (78.1 %)
Tumor grade	Well differentiated	381 (1.2 %)	214 (1.2 %)
	Moderately differentiated	16549 (51.1 %)	7340 (41.7 %)
	Poorly differentiated	15374 (47.5 %)	10024 (56.9 %)
	Undifferentiated	55 (0.2 %)	36 (0.2 %)
Prior Charlson	0	20238 (62.5 %)	11890 (67.5 %)
comorbidity score	1	7067 (21.8 %)	3699 (21.0 %)
	≥ 2	5054 (15.6 %)	2025 (11.5 %)
Year	2004	3076 (9.5 %)	1674 (9.5 %)
	2005	3003 (9.3 %)	1653 (9.4 %)
	2006	3365 (10.4 %)	1879 (10.7 %)
	2007	3419 (10.6 %)	2027 (11.5 %)
	2008	3315 (10.2 %)	1937 (11.0 %)
	2009	3382 (10.5 %)	1843 (10.5 %)
	2010	3315 (10.2 %)	1884 (10.7 %)
	2011	3568 (11.0 %)	1924 (10.9 %)
	2012	2964 (9.2 %)	1430 (8.1 %)
	2013	2952 (9.1 %)	1363 (7.7 %)
Total claims codes	Mean (SD)	44.3 (34.0)	45.9 (31.7)

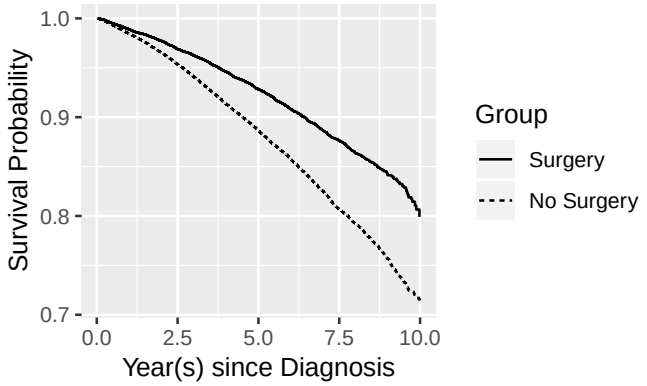
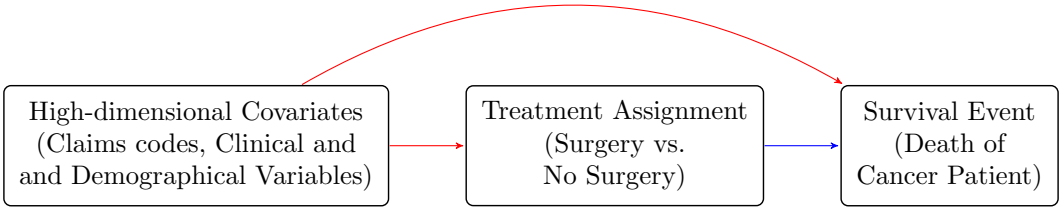
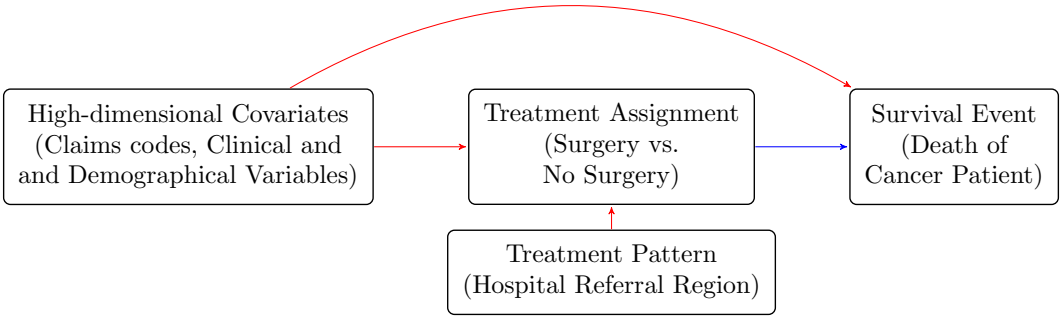


Figure 1: Kaplan-Meier curves for surgery (solid) versus no surgery (dashed).



(a) Analysis I: adjust for the clinical and demographic variables and the claims codes.



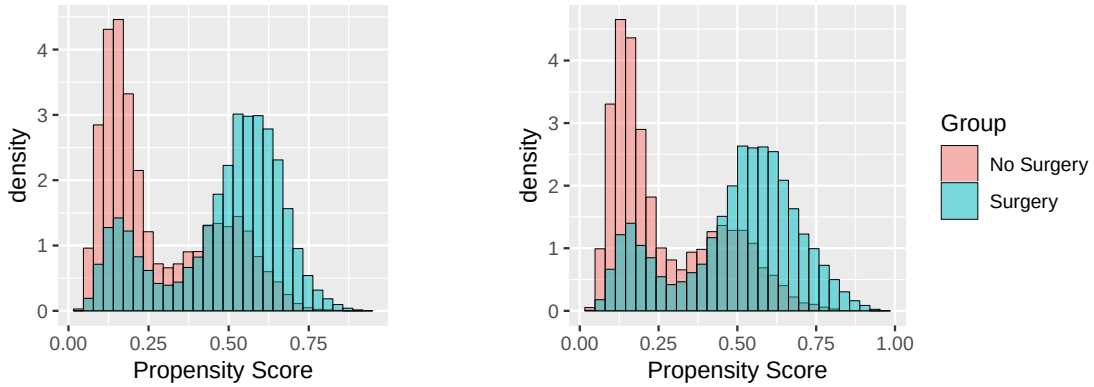
(b) Analysis II: accounting for heterogeneity in treatment pattern as reflected in hospital referral regions (HRR) and its interactions with the covariates from Analysis I in the PS model.

Figure 2: Causal diagrams of the two analyses.

under the treatment propensity score (PS) model selected 309 covariates, and under the additive hazards model selected 378 covariates. In Figure 3, we plot the distribution of the estimated propensity scores from both groups. We note that the range of PS for the surgery group was $0.04 \sim 0.93$, and for the no surgery group was $0.03 \sim 0.90$.

We report the analysis results in Table 5. For comparison purposes, we also present analysis results with only low dimensional covariates: the crude analysis is without adjusting for any covariates, the regression adjustment directly adjusted for the clinical and demographical variables in the additive hazards model, and the inverse probability weighting (IPW) with propensity score estimated using R package '*twang*', and with robust variance estimation. When including the high dimensional covariates, we report the results of the naive additive hazards model Lasso estimate $\tilde{\theta}$ that did not penalize the treatment effect. We also report the results of the IPW estimate with PS estimated by Lasso.

All analysis results suggest that surgery improved overall survival compared to no surgery. The magnitude of the estimated treatment effect, however, varied according to the approach used. The crude analysis had the largest estimated treatment effect of almost 0.01 reduction in hazard rate. Regression adjustment gave an estimated reduction of 0.006, while IPW with low dimensional PS, as well as the four proposed estimators gave reduction around 0.004. It is seen that adjusting for the high dimensional claims codes generally shrank the estimated treatment effects towards zero, perhaps implying that additional confounding had been accounted for. While IPW with Lasso PS had the smallest estimated absolute treatment effect, this approach was not recommended by Belloni et al. (2013) because Lasso gave biased estimate of PS which in turn could not properly account for the true amount of confounding when used to form the weights. We also note that $\check{\theta}_{cf}$ in particular provided a doubly robust estimate requiring only one of the two fitted models to be correct.



(a) Analysis I (b) Analysis II

Figure 3: Distribution of the estimated propensity scores.

Table 5: Estimated treatment effect ($\times 10^{-3}$) from the linked SEER-Medicare data. Crude analysis did not adjust for any covariates. Lasso estimator $\tilde{\theta}$ penalized only the covariate effects β but not θ. $\hat{\theta}$, $\check{\theta}$, $\hat{\theta}_{cf}$ and $\check{\theta}_{cf}$ are the four proposed estimators, where the subscript 'cf' denotes the cross-fitting.

Approach	Estimate	SE	95 % CI	p-value
Low-dimensional analysis: $p = 23$.				
Crude	-9.971	0.605	[-11.157 , -8.786]	< 0.001
Regression adjustment	-6.151	0.722	[-7.567 , -4.735]	< 0.001
IPW with PS	-4.408	0.757	[-5.893 , -2.924]	< 0.001
Analysis I: $p = 6533, \hat{s}_\beta = 378, \hat{s}_\gamma = 309$.				
Lasso $\tilde{\theta}^*$	-5.598	–	–	–
IPW with Lasso PS	-2.603	0.873	[-4.314 , -0.892]	0.003
$\hat{\theta}$	-4.193	0.730	[-5.624 , -2.761]	< 0.001
$\check{\theta}$	-4.187	0.730	[-5.619 , -2.756]	< 0.001
$\hat{\theta}_{cf}$	-3.851	0.733	[-5.288 , -2.414]	< 0.001
$\check{\theta}_{cf}$	-4.310	0.732	[-5.746 , -2.875]	< 0.001
Analysis II: $p_\beta = 6533, \hat{s}_\beta = 378, p_\gamma = 43466, \hat{s}_\gamma = 883$.				
Lasso $\tilde{\theta}^*$	-5.598	–	–	–
IPW with Lasso PS	-2.945	0.863	[-4.636 , -1.253]	0.001
$\hat{\theta}$	-4.151	0.735	[-5.591 , -2.711]	< 0.001
$\check{\theta}$	-4.149	0.735	[-5.589 , -2.708]	< 0.001
$\hat{\theta}_{cf}$	-3.784	0.738	[-5.231 , -2.337]	< 0.001
$\check{\theta}_{cf}$	-4.271	0.738	[-5.717 , -2.826]	< 0.001

* Inference is not directly available. Only the estimates are reported.

In addition to the above, heterogeneity in treatment pattern has been noticed across geographic regions (Harlan et al., 2001). In our data, geographic region is described by the hospital referral region (HRR). In the previous analyses, HRR was used to construct instrumental variables (Hadley et al., 2010; Ying et al., 2019). For high-dimensional data, however, Belloni et al. (2013) recommended that covariates associated with either the treatment or the outcome be included. The goal here, as discussed in Belloni et al. (2013), is to try to model both the treatment and the outcome as closely to the truth as possible.

In Analysis II, we included in the treatment propensity score model all the interactions between HRR and the covariates in Analysis I. After excluding claims codes and binary interaction terms with less than 10 occurances, we have 43466 covariates for the PS model in Analysis II. The large number of additional interaction terms puts the PS model in the $p \approx n$ scenario. As it turned out, the analysis results were numerically stable and quantitatively similar to Analysis I above, despite the dramatically increased number of covariates in the PS model.

6 Discussion

In this paper we have developed the treatment (i.e. propensity score) model in a novel way so that the resulting estimate of the treatment effects with biased input from regularized regression is consistent and asymptotically normal (at root-n rate). In addition, we have provided several refinements to our proposed method to achieve doubly robust estimation in the cases where the propensity score model might be wrong, or the specified survival model might be wrong, or the sparsity assumption is violated. This is achieved via the HDi estimator, which can be seen as a special case of the orthogonal score family, and via cross-fitting (also known as data-splitting), which relaxes the model sparsity condition. Our result on double robustness extends the existing work in that it no longer requires convergence to a defined 'least false' parameter, but instead provides results using simple cross-validation; to the best of our knowledge this is a unique result. We have also developed a novel sparsity doubly robust result where either the outcome or the propensity model can be a fully dense high dimensional model.

Compared to the existing literature on the inference problems with high-dimensional data, our paper has unique contributions. Chernozhukov et al. (2018a) studied the inference on the treatment effect under partially linear conditional mean model using orthogonal score. The proposed inference method in their Section 4 is very general, but it cannot be applied to survival data with censoring. Indeed the commonly used models for survival outcome are conditional hazard models and not the conditional mean models. Our methodology makes substantial contribution to the analysis of censored data based on the martingale techniques under the conditional hazard models. By focusing on the Lasso, we are able to give clear theory for the orthogonal score approach Section 2. In addition, we developed the double robust results in Section 3.

On a separate front, the one-step debiasing methods of Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014) address the inference problem for low-dimension projection of the coefficients from the high-dimensional regression. This can be applied to drawing inference on the treatment effect. The main requirement of these one-step corrections is the consistent estimation of a sparse precision matrix, i.e. the negative inverse Hessian. Compared to our sparsity assumption on the propensity model, the sparsity condition on the precision matrix is harder to interpret or to verify in practice. For finite sample performance, the one-step corrected Lasso has been reported to have substantial under-coverage of the confidence intervals for the non-zero coefficients (Dezeure et al., 2015). In our simulations, the proposed approaches in this paper have shown reasonable coverage for the non-zero treatment effect.

In this paper we have only considered a point (i.e. fixed, as opposed to time-varying) binary treatment.

Although time-dependent covariates can be allowed by the mathematical theory developed here, they pose an issue of causal interpretation when later covariate values are affected by earlier treatment assignment. The treatment effect can be defined through the structural nested models (Robins, 2004; Vansteelandt and Joffe, 2014), a setting that is far beyond the scope of the current work.

The other forms of our orthogonal score method, $\hat{\theta}$ in Section 2, the HDi estimator $\check{\theta}$ without cross-fitting in Section 3.1, and $\hat{\theta}_{cf}$ in Section 3.2, may also produce doubly robust estimation, as suggested by our Lemma 3 and simulation study. The estimators constructed with traditional Breslow estimator require obscure conditions for identifiability when the propensity score model is either wrong or dense, therefore indicating a certain optimality of the weighted Breslow estimator. The estimators without cross-fitting generally require stronger conditions including the uniform consistency under the correct model and the bounded magnitude defined with training data. Here we choose not to expand on those results to avoid unnecessarily lengthy paper.

Acknowledgement

The authors thank our long term collaboration Dr. James Murphy for providing the SEER-Medicare data and many helpful discussions.

Supplementary Materials

In this document we provide details of all of the theoretical results. We present the proofs of the Theorems and Lemmas stated in the main text in Appendix A. The auxiliary results needed in the proofs, including classical and new concentration results, are stated and proved in Appendix B.

A Proof of Main Results

We give the proofs of the Theorems and Lemmas in the order of appearance in the main text.

Proof of Lemma 1. We first verify the identifiability of the true parameters. At the true parameters $(\theta_0, \beta_0, \Lambda_0, \gamma_0)$, $M_i(t; \theta_0, \beta_0, \Lambda_0)$ is a martingale with respect to filtration $\mathcal{F}_{n,t} = \sigma\{N_i(u), Y_i(u), D_i, Z_i : u \leq t, i = 1, \ldots, n\}$. Since the other elements D_i and Z_i are all measurable with respect to $\mathcal{F}_{n,t}$, the martingale integral $\phi(\theta_0; \beta_0, \Lambda_0, \gamma_0)$ is also a $\mathcal{F}_{n,t}$-martingale. Therefore, $\mathbb{E}\{\phi(\theta_0; \beta_0, \Lambda_0, \gamma_0)\} = 0$.

To show the orthogonality, we define the directional perturbations

$$\beta_r = \beta_0 + r\triangle\beta,\ \Lambda_r(t) = \Lambda_0(t) + r\triangle\Lambda(t) \text{ and } \gamma_r = \gamma_0 + r\triangle\gamma.$$

We decompose the expected directional derivative in nuisance parameters evaluated at the true parameters

into 2 terms,

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}\{\phi(\theta_0;\beta_r,\Lambda_r,\gamma_r)\}\Big|_{r=0} \\
=&-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\right\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{\triangle\beta^\top Z_i dt+d\triangle\Lambda(t)\right\}\right] \\
&-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{e^{\gamma_0^\top\tilde{Z}_i}}{\left(1+e^{\gamma_0^\top\tilde{Z}_i}\right)^2}\triangle\gamma^\top\tilde{Z}_i\int_0^\tau e^{D_i\theta_0 t}dM_i(t;\theta_0,\beta_0,\Lambda_0)\right].
\end{aligned}$$

The effect of treatment D_i on the conditional expectation of the at-risk process

$$\mathbb{E}\{Y_i(t)|D_i,Z_i\}=\mathbb{P}(T_i\geq t|D_i,Z_i)\mathbb{P}(C_i\geq t|D_i,Z_i)$$

has two components, the effect on the event-time and that on the censoring time. Under Assumption 1-ii,

$$\mathbb{P}(C_i\geq t|D_i,Z_i)=\mathbb{P}(C_i\geq t|Z_i) \tag{A.1}$$

is $\sigma\{Z_i\}-$measurable. Under model (8),

$$\mathbb{P}(T_i\geq t|D_i,Z_i)=e^{-D_i\theta_0 t-\int_0^t g_0(u;Z_i)du}. \tag{A.2}$$

Therefore, we have the following representation

$$\mathbb{E}[e^{D_i\theta_0 t}Y_i(t)|D_i,Z_i]=\mathbb{P}(C_i\geq t|Z_i)e^{-\int_0^t g_0(u;Z_i)du}=\mathbb{E}\{Y_i(t)|Z_i,D_i=0\}, \tag{A.3}$$

which is obviously $\sigma\{Z_i\}-$measurable. Using the fact $\mathbb{E}\left\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)|Z_i\right\}=0$ under model (3), we apply the tower property of conditional expectation to calculate that the first term equals zero,

$$\int_0^\tau\mathbb{E}\left[\mathbb{E}\left\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)|Z_i\right\}\mathbb{E}\left\{e^{D_i\theta_0 t}Y_i(t)|D_i,Z_i\right\}\left\{\triangle\beta^\top Z_i dt+d\triangle\Lambda(t)\right\}\right]=0.$$

The second term is again a $\mathcal{F}_{n,t}$-martingale, so it also has mean zero. □

Proof of Theorem 1. We use the orthogonality of the score (6) to establish

$$\hat{\theta}-\theta_0=\frac{\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)+o_p(|\hat{\theta}-\theta_0|)}{n^{-1}\sum_{i=1}^n D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0} \tag{A.4}$$

under Assumption 1. The proof of (A.4) involves tedious calculation, so we present the proof separately in Lemma A1.

When the dimension of covariates Z is fixed, the representation (A.4) immediately leads to asymptotic normality through mere formality. However, the growing dimension of covariates in our high-dimensional setting may cause the violation of the classical boundedness assumptions on the summands of $\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)$.

The rest of our proof takes 4 steps. First, we show that $\hat{\theta}$ is consistent for θ_0. Second, we establish the asymptotic normality of the score ϕ at true parameter. Third, we obtain the $\sqrt{n}$-tightness and the asymptotic distribution of $\hat{\theta}-\theta_0$. Finally, we show that the variance estimator is consistent.

Step 1:
Under model (1),

$$\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}e^{\theta_0 D_i t}dM_i(t) \tag{A.5}$$

is a martingale with respect to filtration $\mathcal{F}_{n,t}=\sigma\{N_i(u),Y_i(u),D_i,Z_i:u\le t,i=1,\ldots,n\}$

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^{t}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}e^{\theta_0 D_i t}dM_i(t)$$

evaluated at $t=\tau$. Its expectation is thus zero,

$$\mathbb{E}\{\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)\}=0. \tag{A.6}$$

The true θ_0 is thus identified by the estimating equation $\phi(\theta;\hat{\beta},\hat{\Lambda}(\theta),\hat{\gamma})=0$. From (A.6), we may apply the concentration result of Lemma A9, getting

$$\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0,W_i)=o_p(1). \tag{A.7}$$

Under Assumption 1-ii, we use the martingale property of $M(t)$, defined by (2), and Lemma A13 to calculate the derivative with respect to θ at θ_0

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\mathbb{E}\left\{\phi(\theta;\beta_0,\Lambda_0,\gamma_0)\right\}\Big|_{\theta=\theta_0}\\
=\;&\mathbb{E}\mathbb{E}\left(\{D-\text{expit}(\gamma_0^\top\tilde{Z})\}D\mathbb{E}\left[\int_0^{\tau}e^{D\theta_0 t}\{tdM(t)-Y(t)dt\}\Big|D,Z\right]\Big|Z\right)\\
=\;&-\mathbb{E}\left[D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0\right].
\end{aligned} \tag{A.8}$$

Notice the discontinuity of $(e^{\theta_0 X_i}-1)/\theta_0$ at $\theta_0=0$ can be removed as $\lim_{\theta_0\to 0}(e^{\theta_0 X_i}-1)/\theta_0=X_i$. We have under the logistic regression model

$$\mathbb{E}[D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}|Z_i]=\text{Var}(D_i|Z_i),$$

and under the additive hazards model

$$\mathbb{E}\{(e^{\theta_0 X_i}-1)/\theta_0|Z_i\}=\mathbb{E}\left\{\int_0^{\tau}e^{D_i\theta_0 t}Y_i(t)dt|Z_i\right\}=\mathbb{E}\{Y(t)|Z;D=0\},$$

so we have an alternative representation of (A.8),

$$-\mathbb{E}\left[D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0\right]=-\int_0^{\tau}\mathbb{E}[\mathbb{E}\{Y(t)|Z;D=0\}\text{Var}(D|Z)]dt.$$

Under Assumption 1-v, we use the fact that $Y(t)$ is decreasing in time with minimum at $Y(\tau)$ to bound (A.8) away from zero

$$\int_0^{\tau}\mathbb{E}[\mathbb{E}\{Y(t)|Z;D=0\}\text{Var}(D|Z)]dt\ge\int_0^{\tau}\mathbb{E}[\mathbb{E}\{Y(\tau)|Z;D=0\}\text{Var}(D|Z)]dt\ge\tau\varepsilon_Y. \tag{A.9}$$

Since the summands in the denominator of (A.4) have bound

$$|D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0| \le e^{\tau\theta_0}\tau, \tag{A.10}$$

we can use the Hoeffding's inequality (as in Lemma A3) to establish a lower bound

$$\mathbb{P}\left(n^{-1}\sum_{i=1}^n D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0 > \varepsilon_Y/2\right) > 1-e^{-\frac{n\varepsilon_Y^2}{8e^{2\tau\theta_0}\tau^2}}. \tag{A.11}$$

Plugging the rate (A.7) and the lower bound (A.11) into (A.4), we conclude that $\hat{\theta}-\theta_0 = o_p(1)$.

Step 2: Let $X_{(1)},\ldots,X_{(n)}$ be the order statistics of the observed times and

$$\begin{aligned} M_k^1 &= \frac{1}{n}\sum_{i=1}^n \int_0^{X_{(k)}} D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{\theta_0 t}dM_i(t), \\ M_k^0 &= \frac{1}{n}\sum_{i=1}^n \int_0^{X_{(k)}} (1-D_i)\text{expit}(\gamma_0^\top \tilde{Z}_i)dM_i(t), \end{aligned} \tag{A.12}$$

for $k=0,\ldots,n$. We note that the score ϕ with true parameters can be alternatively expressed as

$$\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0) = M_n^1 - M_n^0. \tag{A.13}$$

Since the integrands of both M_k^1 and M_k^0, $D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{\theta_0 t}$ and $(1-D_i)\text{expit}(\gamma_0^\top \tilde{Z}_i)$, in (A.12) are nonnegative and bounded by $\tau(1\vee e^{\theta_0\tau})$, we can apply the Lemma A8 to get that both M_k^1 and M_k^0, hence $M_k^1 - M_k^0$, are martingales under filtration $\mathcal{F}_k^M = \sigma\{N_i(u), Y_i(u+), D_i, Z_i : u\in[0,t_k], i=1,\ldots,n\}$ satisfying, most importantly,

$$\max\left\{\mathbb{E}\left\{(M_k^1-M_{k-1}^1)^2|\mathcal{F}_k^M\right\}, \mathbb{E}\left\{(M_k^0-M_{k-1}^0)^2|\mathcal{F}_k^M\right\}\right\} \le 8\tau^2(1\vee e^{\theta_0\tau})^2/n^2. \tag{A.14}$$

By the Cauchy-Schwartz inequality, we have

$$(M_k^1 - M_k^0 - M_{k-1}^1 + M_{k-1}^0)^2 \le 2(M_k^1-M_{k-1}^1)^2 + 2(M_k^0-M_{k-1}^0)^2. \tag{A.15}$$

Hence for the quadratic variation of $\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)$, we have

$$\mathbb{E}\left\{(M_k^1-M_{k-1}^1-M_k^0+M_{k-1}^0)^2|\mathcal{F}_k^M\right\} \le 32\tau^2(1\vee e^{\theta_0\tau})^2/n^2. \tag{A.16}$$

As a result, the variance

$$\sigma_\phi^2 = \text{Var}\{\sqrt{n}\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)\} = n\mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left\{(M_k^1-M_k^0-M_{k-1}^1+M_{k-1}^0)^2|\mathcal{F}_k^M\right\}\right] \tag{A.17}$$

is finite, bounded by $32\tau^2(1\vee e^{\theta_0\tau})^2$.

Now, we verify the Lindeberg condition for the martingale central limit theorem (Brown, 1971). The event

$$\sqrt{n}|M_k^1 - M_{k-1}^1 - M_k^0 + M_{k-1}^0| > \varepsilon \tag{A.18}$$

occurs only if

$$\sqrt{n}|M_k^1 - M_{k-1}^1| > \varepsilon/2 \text{ or } \sqrt{n}|M_k^0 - M_{k-1}^0| > \varepsilon/2 \tag{A.19}$$

occurs. Let $I(\cdot)$ be the binary event indicator. Thus, we must have the following inequality

$$\begin{aligned} & I(\sqrt{n}|M_k^1 - M_{k-1}^1 - M_k^0 + M_{k-1}^0| > \varepsilon) \\ \leq \; & I(\sqrt{n}|M_k^1 - M_{k-1}^1| > \varepsilon/2) + I(\sqrt{n}|M_k^0 - M_{k-1}^0| > \varepsilon/2). \end{aligned} \tag{A.20}$$

Along with (A.15), we have

$$\begin{aligned} & n\sum_{i=1}^{n} \mathbb{E}\left\{(M_k^1 - M_{k-1}^1 - M_k^0 + M_{k-1}^0)^2 I\left(\sqrt{n}|M_k^1 - M_k^0 - M_{k-1}^1 + M_{k-1}^0| > \varepsilon\right)\right\} \\ \leq \; & 2n\sum_{i=1}^{n} \mathbb{E}\left\{(M_k^1 - M_{k-1}^1)^2 I\left(\sqrt{n}|M_k^1 - M_{k-1}^1| > \varepsilon/2\right)\right\} \\ & +2n\sum_{i=1}^{n} \mathbb{E}\left\{(M_k^0 - M_{k-1}^0)^2 I\left(\sqrt{n}|M_k^0 - M_{k-1}^0| > \varepsilon/2\right)\right\}. \end{aligned} \tag{A.21}$$

By Lemma A8, the right hand side in (A.21) decays to zero when n approaches ∞. Hence, we can apply the martingale central limit theorem to

$$\sqrt{n}\sigma_\phi^{-1}\phi(\theta_0; \beta_0, \Lambda_0, \gamma_0, W_i) \rightsquigarrow N(0, 1). \tag{A.22}$$

Step 3: We define the asymptotic standard deviation of $\sqrt{n}(\hat{\theta} - \theta_0)$ as

$$\sigma = \sigma_\phi / \mathbb{E}[D\{1 - \text{expit}(\gamma_0^\top Z_1)\}(e^{\theta_0 X} - 1)/\theta_0], \tag{A.23}$$

where σ_ϕ is the square root of (A.17). Since $\hat{\theta}$ solves $\phi(\theta; \hat{\beta}, \hat{\Lambda}(\theta), \hat{\gamma}) = 0$, we have along with Lemma A1

$$\begin{aligned} & \sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0) - \frac{\sqrt{n}}{\sigma_\phi}\phi(\theta_0; \beta_0, \gamma_0, W_i) \\ = \; & \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma_\phi}\Bigg(\mathbb{E}[D\{1 - \text{expit}(\gamma_0^\top Z_1)\}(e^{\theta_0 X} - 1)/\theta_0] \\ & -\frac{1}{n}\sum_{i=1}^{n} D_i\{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}(e^{\theta_0 X_i} - 1)/\theta_0\Bigg) \\ & +o_p(1 + \sqrt{n}|\hat{\theta} - \theta_0|). \end{aligned} \tag{A.24}$$

Again using the bound (A.10), we apply the Hoeffding's inequality (as in Lemma A3) to establish that

$$\mathbb{E}[D\{1 - \text{expit}(\gamma_0^\top Z_1)\}(e^{\theta_0 X} - 1)/\theta_0] - \frac{1}{n}\sum_{i=1}^{n} D_i\{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}(e^{\theta_0 X_i} - 1)/\theta_0 \tag{A.25}$$

is of order $O_p(n^{-1/2})$. Hence, the right hand side of (A.24) is of order $o_p(1 + \sqrt{n}|\hat{\theta} - \theta_0|)$. Along with the normality (A.22), we establish the $\sqrt{n}$-tightness of the estimation error

$$|\hat{\theta} - \theta_0| = O_p\big(n^{-1/2}\big). \tag{A.26}$$

Plugging in the rate of estimation error into the righthand side of (A.24), we obtain the asymptotic equivalence

$$\sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0) - \sqrt{n}\sigma_\phi^{-1}\phi_n(\theta_0; \hat{\beta}, \hat{\gamma}, W) = o_p(1). \tag{A.27}$$

Step 4: To show that $\hat{\sigma}^{-1}$ defined

$$\hat{\sigma}^2 = \frac{n^{-1}\sum_{i=1}^n \delta_i\{D_i - \text{expit}(\hat{\gamma}^\top \tilde{Z}_i)\}^2 e^{2\hat{\theta} D_i X_i}}{\left\{n^{-1}\sum_{i=1}^n (1-D_i)\text{expit}(\hat{\gamma}^\top \tilde{Z}_i) X_i\right\}^2}.$$

is a consistent estimator for σ^{-1}, we decompose the numerator of $\hat{\sigma}^2$ into

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \delta_i\{D_i - \text{expit}(\hat{\gamma}^\top \tilde{Z}_i)\}^2 e^{2\hat{\theta} D_i X_i} \\
= \;& \frac{1}{n}\sum_{i=1}^n \left[\delta_i\{D_i - \text{expit}(\hat{\gamma}^\top \tilde{Z}_i)\}^2 e^{2\hat{\theta} D_i X_i} - \delta_i\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2 e^{2\theta_0 D_i X_i}\right] \\
&+ \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{\theta_0 D_i t}]^2 dN_i(t). \qquad \text{(A.28)}
\end{aligned}$$

By the mean value theorem, the first term in the righthand side of (A.28) can be written in terms of $\theta_\xi = (1-\xi)\theta_0 + \xi\hat{\theta}$ and $\gamma_\xi = (1-\xi)\gamma_0 + \xi\hat{\gamma}$ with some $\xi \in (0,1)$,

$$\begin{aligned}
&\frac{(\gamma_0 - \hat{\gamma})^\top}{n}\sum_{i=1}^n \frac{\delta_i\{D_i - \text{expit}(\gamma_\xi^\top \tilde{Z}_i)\} e^{\gamma_\xi^\top \tilde{Z}_i} e^{2\theta_\xi D_i X_i}}{\left(1 + e^{\gamma_\xi^\top \tilde{Z}_i}\right)^2} \\
&+ \frac{(\hat{\theta} - \theta_0)}{n}\sum_{i=1}^n 2\delta_i D_i X_i \{1 - \text{expit}(\gamma_\xi^\top \tilde{Z}_i)\}^2 e^{2\theta_\xi D_i X_i} \\
= \;& O_p(\|\hat{\gamma} - \gamma_0\|_1 + |\hat{\theta} - \theta_0|). \qquad \text{(A.29)}
\end{aligned}$$

We repeatedly use ξ in all mean value theorem expansions for convenience of notation. The second term on the righthand side of (A.28) is the optional quadratic variation of $\phi(\theta_0; \beta_0, \gamma_0, \Lambda_0)$ bounded by $e^{2\theta_0\tau}$ (Kalbfleisch and Prentice, 2002, (5.17) p. 159 and (5.26) p. 162). By the Hoeffding's inequality (as in Lemma A3), we have the concentration of the second term around the variance of $\phi(\theta_0; \beta_0, \gamma_0, \Lambda_0)$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \int_0^\tau [\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{\theta_0 D_i t}]^2 dN_i(t) \\
= \;& \mathbb{E}\left(\int_0^\tau [\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{\theta_0 D_i t}]^2 dN_i(t)\right) + O_p(n^{-1/2}) \\
= \;& \sigma_\phi^2 + o_p(1). \qquad \text{(A.30)}
\end{aligned}$$

Putting (A.29) and (A.30) together, we have the numerator of $\hat{\sigma}^2$ (A.7) equals $\sigma_\phi^2 + o_p(1)$. Similarly, we decompose the denominator of σ into

$$\frac{1}{n}\sum_{i=1}^n \left[D_i\{1 - \text{expit}(\hat{\gamma}^\top \tilde{Z}_i)\}\frac{e^{\hat{\theta} X_i} - 1}{\hat{\theta}_0} - D_i\{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}\frac{e^{\theta_0 X_i} - 1}{\theta_0}\right] \qquad \text{(A.31)}$$

minus (A.25). Again, we have (A.31) is of order $O_p(|\hat{\theta} - \theta_0| + \|\hat{\gamma} - \gamma_0\|_1) = o_p(1)$ through the mean value theorem. Under the additive hazards model (1), we must have a nonnegative hazard among the control subjects

$$\beta_0^\top Z + d\Lambda_0(t) \geq 0 \qquad \text{(A.32)}$$

for all Z such that $\Pr(D=0|Z)>0$. Under Assumptions 1-ii, 1-v and 1-vi, we can establish a lower bound for σ_ϕ

$$\begin{aligned}
\sigma_\phi &= \mathbb{E}\left[\int_0^\tau \{D-\text{expit}(\gamma_0^\top Z)\}^2 e^{2D\theta_0 t} Y(t)\{(D\theta_0+\beta_0^\top Z)dt+d\Lambda_0(t)\}\right] \\
&= \mathbb{E}\left[\int_0^\tau \{D-\text{expit}(\gamma_0^\top Z)\}^2 e^{D\theta_0 t} \mathbb{E}\{Y(t)|Z;D=0\}\{(D\theta_0+\beta_0^\top Z)dt+d\Lambda_0(t)\}\right] \\
&= \mathbb{E}\left[\int_0^\tau D\{1-\text{expit}(\gamma_0^\top Z)\}^2 e^{\theta_0 t} \mathbb{E}\{Y(\tau)|Z;D=0\}\theta_0 dt\right] \\
&\quad +\mathbb{E}\left[\int_0^\tau \{D-\text{expit}(\gamma_0^\top Z)\}^2 e^{D\theta_0 t} d\mathbb{E}\{N(t)|Z;D=0\}\right] \\
&\geq 0+e^{1\wedge\theta_0\tau}\mathbb{E}[\text{Var}(D|Z)\mathbb{E}\{N(\tau)|Z;D=0\}] \\
&\geq e^{1\wedge\theta_0\tau}\varepsilon_N. \qquad \text{(A.33)}
\end{aligned}$$

Hence, the limit is bounded by

$$\sigma^{-1}=\frac{\mathbb{E}\left[D\{1-\text{expit}(\gamma_0^\top Z_1)\}\frac{e^{\theta_0 X}-1}{\theta_0}\right]}{\sqrt{\mathbb{E}[\delta\{D-\text{expit}(\gamma_0^\top Z_1)\}^2 e^{2D\theta_0 X}]}} \leq \frac{\tau e^{\theta_0\tau}}{\sqrt{e^{1\wedge\theta_0\tau}\varepsilon_N}}. \qquad \text{(A.34)}$$

Therefore, we have

$$\hat{\sigma}^{-1}=\sigma^{-1}+o_p(1) \qquad \text{(A.35)}$$

by continuous mapping theorem.

Combining the results (A.22), (A.27) and (A.35), we obtain

$$\sqrt{n}\hat{\sigma}^{-1}(\hat{\theta}-\theta_0) \rightsquigarrow N(0,1). \qquad \text{(A.36)}$$

□

Proof of Theorem 3. We obtain from Lemma A2 the same representation as (A.4),

$$\hat{\theta}_{cf}-\theta_0=\frac{\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)+o_p(|\hat{\theta}-\theta_0|)}{n^{-1}\sum_{i=1}^n D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0}. \qquad \text{(A.37)}$$

The rest of the proof is identical to the Steps 1-4 in the proof of Theorem 1. □

Proof of Lemma 4. To see that zero is always in the LASSO regularization path, we shall spell out the associated tuning parameter. The gradient of the loss $l_\beta(\beta)=\beta^\top H_n\beta-2\beta^\top h_n$ in the additive hazards model LASSO (11) at $\beta=0$ is

$$\nabla l_\beta(0)=-h_n=-\frac{2}{n}\sum_{i=1}^n\int_0^\tau\left[\left\{Z_i-\bar{Z}(t)\right\}dN_i(t)\right].$$

For any penalty large enough, $\lambda>\|\nabla l_\beta(0)\|_\infty$, we have that $\beta=0$ satisfies the LASSO KKT condition $\|\nabla l_\beta(0)\|_\infty<\lambda$ (Gorst-Rasmussen and Scheike, 2012). Therefore, zero is an element in the regularization path. By the optimality of $\hat{\lambda}_\beta$ according to (25), $l_\beta^*(0)=0$ must be an upper bound for $l_\beta^*(\hat{\beta}(\hat{\lambda}_\beta))$.

Then, we derive the lower bound of $l^*_\beta(\beta)$ related to the magnitude $\mathcal{M}_\beta(\beta)$. We apply the Cauchy-Schwartz inequality to the linear term in $l^*_\beta(\beta)$,

$$\begin{aligned} l^*_\beta(\beta) =& \mathcal{M}_\beta(\beta)^2 - 2\int_0^\tau \mathbb{E}_*\left[\beta^\top \{Z_* - \mu(t)\}\, dN_*(t)\right] \\ =& \mathcal{M}_\beta(\beta)^2 - 2\int_0^\tau \mathbb{E}_*\left[\beta^\top \{Z_* - \mu(t)\}\, Y_*(t) g(t, Z_*) dt\right] \\ \geq& \mathcal{M}_\beta(\beta)\left(\mathcal{M}_\beta(\beta) - 2\sqrt{\int_0^\tau \mathbb{E}_*\left[Y_*(t)g(t,Z_*)^2\right] dt}\right). \end{aligned}$$

Putting the upper bound and lower bound to gather, we must have

$$\mathcal{M}_\beta\left(\hat{\beta}(\hat{\lambda}_\beta)\right) \leq 2\sqrt{\int_0^\tau \mathbb{E}_*\left[Y_*(t)g(t,Z_*)^2\right] dt}.$$

□

Proof of Lemma 5. To see that the intercept only estimator $\hat{\gamma}_0$ is always in the LASSO regularization path, we shall spell out the associated tuning paramter. The intercept only estimator $\hat{\gamma}_0$ makes constant predictions $\text{expit}(\hat{\gamma}_0^\top z) = \bar{D} = \sum_{i=1}^n D_i/n$. The gradient of the loss $\nabla l_\gamma(\gamma) = -n^{-1}\sum_{i=1}^n \{D_i \gamma^\top Z_i - \log(1 - e^{\gamma^\top Z_i})\}$ in the logistic regression LASSO (12) is

$$\nabla l_\gamma(\hat{\gamma}_0) = -\frac{1}{n}\sum_{i=1}^n (D_i - \bar{D}) \begin{pmatrix} 1 \\ Z_i \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{n}\sum_{i=1}^n (D_i - \bar{D}) Z_i \end{pmatrix}.$$

Notice that we follow Friedman et al. (2010) in (12) by leaving the intercept term not penalized. For any penalty large enough, $\lambda > \|\frac{1}{n}\sum_{i=1}^n \{D_i - \bar{D}\} Z_i\|_\infty$, we have the first coordinate in $|\nabla l_\gamma(\hat{\gamma}_0)|$ being zero and the rest strictly smaller than λ. Therefore, $\hat{\gamma}_0$ is an element in the regularization path (Friedman et al., 2010). By the Markov inequality, $\hat{\gamma}_0$ converges to $(\log(1 - 1/\mathbb{E}_*(D_*)), 0, \ldots, 0)$. Under Assumption 1-v, $\varepsilon_Y \leq \mathbb{E}_*(D_*) \leq 1 - \varepsilon_Y$, so we have an upper bound for $l^*_\gamma(\hat{\gamma}_0)$,

$$l^*_\gamma(\hat{\gamma}_0) \leq -\mathbb{E}_*(D_*)\log(\mathbb{E}_*(D_*)) - \{1 - \mathbb{E}_*(D_*)\}\log(1 - \mathbb{E}_*(D_*)) + o_p(1) \leq -\log(\varepsilon_Y) + o_p(1).$$

By the optimality of $\hat{\lambda}_\gamma$ according to (25), the upper bound of $l^*_\gamma(\hat{\gamma}_0)$ must also be an upper bound for $l^*_\gamma(\hat{\gamma}(\hat{\lambda}_\gamma))$.

Define the set $\mathcal{Z} = \{z : \mathbb{E}_*(D_*|Z_* = z) \geq \varepsilon_Y/2, \mathbb{E}_*(1 - D_*|Z_* = z) \geq \varepsilon_Y/2$ and $\mathbb{E}_*(Y_*(\tau)|Z_* = z, D_* = 0) \geq \varepsilon_Y/2\}$. We decompose

$$\begin{aligned} & \mathbb{E}_*[\text{Var}_*(D_*|Z_*)\mathbb{E}_*\{Y_*(\tau)|Z_*, D_* = 0\}] \\ =& \mathbb{E}_*[\mathbb{E}_*(D_*|Z_*)\mathbb{E}_*(1 - D_*|Z_*)\mathbb{E}_*\{Y_*(\tau)|Z_*, D_* = 0\}] \\ =& \mathbb{E}_*[\mathbb{E}_*(D_*|Z_*)\mathbb{E}_*(1 - D_*|Z_*)\mathbb{E}_*\{Y_*(\tau)|Z_*, D_* = 0\} I(Z_* \in \mathcal{Z})] \\ & + \mathbb{E}_*[\mathbb{E}_*(D_*|Z_*)\mathbb{E}_*(1 - D_*|Z_*)\mathbb{E}_*\{Y_*(\tau)|Z_*, D_* = 0\} I(Z_* \in \mathcal{Z}^c)] \\ \leq& \mathbb{P}_*(Z_* \in \mathcal{Z}) + \varepsilon_Y/2. \end{aligned}$$

To satisfy Assumption 1-v, $\mathbb{P}_*(Z_* \in \mathcal{Z})$ must be at least $\varepsilon_Y/2$. Then, we derive a lower bound of $l^*_\gamma(\gamma)$ by analyzing the expectation in set $\mathcal{Z}$

$$\begin{aligned}
l^*_\gamma(\gamma) =& -\mathbb{E}_*[D_* \log\{\text{expit}(\gamma^\top Z_*)\} + (1-D_*)\log\{1-\text{expit}(\gamma^\top Z_*)\}] \\
\geq& -\mathbb{E}_*[D_* \log\{\text{expit}(\gamma^\top Z_*)\} + (1-D_*)\log\{1-\text{expit}(\gamma^\top Z_*)\}|Z_* \in \mathcal{Z}]\mathbb{P}_*(Z_* \in \mathcal{Z}) \\
\geq& -\varepsilon_Y^2/4\mathbb{E}_*[\log\{\text{expit}(\gamma^\top Z_*)\}|Z_* \in \mathcal{Z}] - \varepsilon_Y^2/4\mathbb{E}_*[\log\{1-\text{expit}(\gamma^\top Z_*)\}|Z_* \in \mathcal{Z}] \\
\geq& -\varepsilon_Y^2/4\log\left(\mathbb{E}_*\{\text{expit}(\gamma^\top Z_*)|Z_* \in \mathcal{Z}\}\right) - \varepsilon_Y^2/4\log\left(\mathbb{E}_*\{1-\text{expit}(\gamma^\top Z_*)|Z_* \in \mathcal{Z}\}\right).
\end{aligned}$$

The last step above is the consequence of the Jensen's inequality.

Putting the upper bound and lower bound of $l^*_\gamma(\hat{\gamma}(\hat{\lambda}_\gamma))$ together, we have

$$\begin{aligned}
&\mathbb{E}_*\{\text{expit}(\hat{\gamma}(\hat{\lambda}_\gamma)^\top Z_*)|Z_* \in \mathcal{Z}\} \geq e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}, \\
&\mathbb{E}_*\{1-\text{expit}(\hat{\gamma}(\hat{\lambda}_\gamma)^\top Z_*)|Z_* \in \mathcal{Z}\} \geq e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}.
\end{aligned}$$

The bounds above are connected to $\mathcal{M}_\gamma\left(\hat{\gamma}(\hat{\lambda}_\gamma)\right)$ through

$$\begin{aligned}
&\mathbb{E}_*\{w^0_*(\hat{\gamma}(\hat{\lambda}_\gamma))X_*\} \geq \tau\varepsilon_Y^3/8\mathbb{E}_*\{\text{expit}(\hat{\gamma}(\hat{\lambda}_\gamma)^\top Z_*)|Z_* \in \mathcal{Z}\}, \\
&\mathbb{E}_*\{w^1_*(\hat{\gamma}(\hat{\lambda}_\gamma))Y_*(\tau)\} \geq \varepsilon_Y^3/8\mathbb{E}_*\{1-\text{expit}(\hat{\gamma}(\hat{\lambda}_\gamma)^\top Z_*)|Z_* \in \mathcal{Z}\}.
\end{aligned}$$

Therefore, we obtain the bound $\mathcal{M}_\gamma\left(\hat{\gamma}(\hat{\lambda}_\gamma)\right) \leq (1+\tau^{-1})8\varepsilon_Y^{-3}e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}$. □

Proof of Theorem 4. We prove the theorem under two setups given by Assumptions 3(a) and 3(b) separately. We denote the cross-fitted weighted Breslow estimator $\check{\Lambda}$ defined in (17) as

$$\check{\Lambda}^{(j)}(t,\theta;\beta,\gamma) = \int_0^t \frac{\sum_{i\in I_j} w^1_i(\gamma)\{dN_i(u) - Y_i(u)(D_i\theta + \beta^\top Z_i)du\}}{\sum_{i\in I_j} w^1_i(\gamma)Y_i(u)}, \tag{A.38}$$

constructed with samples in fold-j. We denote the cross-fitted score associated with the closed form estimator $\check{\theta}_{cf}$ for fold-j as

$$\begin{aligned}
&\psi^{(j)}(\theta;\beta,\gamma) \\
=&\phi^{(j)}\left(\theta;\beta,\check{\Lambda}^{(j)}(\cdot,\theta;\beta,\gamma),\gamma\right) \\
=&-\frac{1}{n}\sum_{i\in I_j} w^0_i(\hat{\gamma}^{(j)})\int_0^\tau \left(dN_i(u) - Y_i(u)\left[\hat{\beta}^{(j)\top}\{Z_i - \check{Z}^{(j)}(u;\hat{\gamma}^{(j)})\}du + d\widetilde{N}^{(j)}(u;\hat{\gamma}^{(j)})\right]\right) \\
&-\frac{\theta}{n}\sum_{i\in I_j}(1-D_i)\text{expit}(\gamma^\top\tilde{Z}_i)X_i,
\end{aligned} \tag{A.39}$$

(a) First, we show that the true parameter is identified by the score. That is

$$\psi^{(j)}\left(\theta_0;\hat{\beta}^{(j)},\hat{\gamma}^{(j)}\right) = o_p(1). \tag{A.40}$$

We decompose

$$
\begin{aligned}
& \psi^{(j)}\left(\theta_0; \hat{\beta}^{(j)}, \hat{\gamma}^{(j)}\right) \\
= & -\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt \\
& +\frac{1}{n}\sum_{i'\in I_j}\int_0^\tau \frac{\sum_{i\in I_j}\{D_i - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)}{\sum_{i\in I_j} w_i^1(\hat{\gamma}^{(j)})Y_i(t)} w_{i'}^1(\hat{\gamma}^{(j)})Y_{i'}(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_{i'}dt \\
& +\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\hat{\gamma}^{(j)}\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}dM_i(t) \\
& -\frac{1}{n}\sum_{i'\in I_j}\int_0^\tau \frac{\sum_{i\in I_j}\{D_i - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)}{\sum_{i\in I_j} w_i^1(\hat{\gamma}^{(j)})Y_i(t)} w_{i'}^1(\hat{\gamma}^{(j)})dM_{i'}(t) \\
= & \ Q_1 + Q_2 + Q_3 + Q_4. \qquad \text{(A.41)}
\end{aligned}
$$

We shall show that each term Q_1-Q_4 in (A.41) is negligible.

By applying twice the Cauchy-Schwartz inequality, first to the sum then to the integral, we have a bound for Q_1,

$$
|Q_1| \le \frac{1}{n}\sum_{i\in I_j} 1 e^{K_\theta \tau}\int_0^\tau Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt \le \frac{1}{n}\sqrt{|I_j|e^{K_\theta\tau}}\sqrt{\sum_{i\in I_j}\{(\hat{\beta}^{(j)}-\beta_0)^\top Z_i\}^2 X_i}. \qquad \text{(A.42)}
$$

Also from Assumption 3a-ii, the squared average testing deviance $\mathbb{E}_*\{(\hat{\beta}^{(j)}-\beta_0)^\top Z_*\}^2 X_*$ converges to zero. Applying the Markov inequality conditioning on the out-of-fold data, we have its asymptotic equivalence to the empirical counterpart

$$
\frac{1}{|I_j|}\sum_{i\in I_j}\{(\hat{\beta}^{(j)}-\beta_0)^\top Z_i\}^2 X_i = \mathbb{E}_*\{(\hat{\beta}^{(j)}-\beta_0)^\top Z_*\}^2 X_* + o_p(1) = o_p(1). \qquad \text{(A.43)}
$$

Plugging (A.43) to (A.42), we conclude that $Q_1 = o_p(1)$.

Similarly for Q_2, we apply the Cauchy-Schwartz inequality twice,

$$
\begin{aligned}
|Q_2| \le & \frac{1}{n}\sqrt{\int_0^\tau \left[\frac{\sum_{i\in I_j}\{D_i - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)}{\sum_{i\in I_j} w_i^1(\hat{\gamma}^{(j)})Y_i(t)}\right]^2 \sum_{i\in I_j} w_i^2(t;\theta_0,\hat{\gamma}^{(j)})Y_i(t)dt} \\
& \times \sqrt{\sum_{i\in I_j}\{(\hat{\beta}^{(j)}-\beta_0)^\top Z_i\}^2 X_i}. \qquad \text{(A.44)}
\end{aligned}
$$

From Assumption 3a-ii, we have a lower bound for $\mathbb{E}_*\{w_*^1(\hat{\gamma}^{(j)})Y_*(\tau)\} \ge K_{\mathcal{M}_g}^{-1}$. Applying the Hoeffding's inequality to the empirical version of the process, we get $\frac{1}{|I_j|}\sum_{i\in I_j} w_i^1(\hat{\gamma}^{(j)})Y_i(\tau) \ge K_{\mathcal{M}_g}^{-1}/2$ with probability tending to one. The denominator term in Q_2 is decreasing process in t thus achieves it minimal at $t=\tau$, so it has the lower bound $K_{\mathcal{M}_g}^{-1}/2$ with probability tending to one. Along with (A.43), we conclude from (A.44) with probability tending to one

$$
Q_2 \le 2e^{2K_\theta\tau}\tau K_{\mathcal{M}_g} O_p\left(\mathcal{D}_{\beta *}\left(\hat{\beta}^{(j)},\beta_0\right)\right) = o_p(1). \qquad \text{(A.45)}
$$

Q_3 and Q_4 are martingale integrals with respect to filtration

$$\mathcal{F}_{I_j,t} = \sigma\left(\{(N_i(u), Y_i(u), D_i, Z_i) : i \in I_j, u \leq t\} \cup \{(X_i, \delta_i, D_i, Z_i) : i \in I_{-j}\}\right).$$

The integrands are bounded with probability tending to one under Assumption 3a-ii, so we obtain by Lemma A9 that $Q_3 = O_p\left(n^{-1/2}\right)$ and $Q_4 = O_p\left(n^{-1/2}\right)$.

We combine the results for $Q_1 - Q_4$ to establish the identifiability result (A.40).

By the Assumption 3a-ii, we have the denominator in $\check{\theta}$ (27)

$$Q' = -\frac{1}{n}\sum_{j=1}^{k}\sum_{i\in I_j}(1 - D_i)\text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_j)X_i. \tag{A.46}$$

bounded from below by $2kK_{\mathcal{M}_g}^{-1}$ with probability tending to one.

Utilizing the linearity of ψ, we can write

$$(\check{\theta} - \theta_0) = \{Q'\}^{-1}\sum_{j=1}^{k}\psi^{(j)}(\theta_0; \hat{\beta}^{(j)}, \hat{\gamma}^{(j)}) = o_p(1). \tag{A.47}$$

We hence obtain the consistency of $\check{\theta}$.

(b) Under model (8), we have for $i \in I_j$ the following martingale with respect to filtration $\mathcal{F}_{I_j,t} = \sigma\left(\{(N_i(u), Y_i(u), D_i, Z_i) : i \in I_j, u \leq t\} \cup \{(X_i, \delta_i, D_i, Z_i) : i \in I_{-j}\}\right)$

$$M_i(t) = N_i(t) - \int_0^{\tau} Y_i(t)\{D_i\theta_0 + g_0(t; Z_i)\}dt. \tag{A.48}$$

First, we prove the identifiability result like (A.40). We decompose

$$
\begin{aligned}
& \psi^{(j)}(\theta_0; \hat{\beta}^{(j)}, \hat{\gamma}^{(j)}) \\
= \; & \frac{1}{n}\sum_{i\in I_j}\{D_i - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}dM_i(t) \\
& + \frac{1}{n}\sum_{i\in I_j}\int_0^\tau \{D_i - \mathrm{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)g_0(t;Z_i)dt \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{D_i - \mathrm{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}dt \\
& + \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{D_i - \mathrm{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\{\theta_0 dt - d\widetilde{N}^{(j)}(t;\hat{\gamma}^{(j)})\} \\
& + \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\mathrm{expit}(\gamma_0^\top \tilde{Z}_i) - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)g_0(t;Z_i)dt \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\mathrm{expit}(\gamma_0^\top \tilde{Z}_i) - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}dt \\
& + \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\mathrm{expit}(\gamma_0^\top \tilde{Z}_i) - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\{\theta_0 dt - d\widetilde{N}^{(j)}(t;\hat{\gamma}^{(j)})\} \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{D_i - \mathrm{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\hat{\beta}^{(j)\top}\{\mu(t) - \check{Z}^{(j)}(t;\hat{\gamma}^{(j)})\}dt \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\mathrm{expit}(\gamma_0^\top \tilde{Z}_i) - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\hat{\beta}^{(j)\top}\{\mu(t) - \check{Z}^{(j)}(t;\hat{\gamma}^{(j)})\}dt \\
= \; & Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6 + Q_7 + Q_8 + Q_9. \qquad \text{(A.49)}
\end{aligned}
$$

Q_1 is the final element of the $\mathcal{F}_{I_j,t}$-martingale,

$$
Q_{1,t} = \frac{1}{n}\sum_{i\in I_j}\{D_i - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}\int_0^t e^{D_i\theta_0 u}dM_i(u). \qquad \text{(A.50)}
$$

The measurable quadratic variation of $Q_{1,t}$ is

$$
\langle Q_{1,\cdot}\rangle_t = \frac{1}{n^2}\sum_{i\in I_j}\{D_i - \mathrm{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}^2\int_0^t e^{2D_i\theta_0 u}Y_i(u)g_0(t;Z_i)du. \qquad \text{(A.51)}
$$

By the Cauchy-Schwartz's inequality, we have the upper bound for

$$
\mathrm{Var}(Q_1) = \mathbb{E}\langle Q_{1,\cdot}\rangle_\tau \le \mathbb{E}\left\{\frac{1}{n^2}\sqrt{ne^{2K_\theta\tau}\tau}\sqrt{\sum_{i\in I_j}\int_0^\tau g_0(t;Z_i)dt}\right\}.
$$

Under Assumption 3(b), we have

$$
\sqrt{\sum_{i\in I_j}\int_0^\tau g_0(t;Z_i)dt} = O_p(nK_\Lambda)
$$

by Markov's inequality. Thus, we have $\text{Var}(Q_1) = O(K_\Lambda/n) = o(1)$. By the Tchebychev's inequality, we obtain $Q_1 = o_p(1)$.

Using Lemma A13, we have for Q_2

$$\mathbb{E}[\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)g_0(t;Z_i)] = \mathbb{E}(\mathbb{E}[\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)|Z_i]g_0(t;Z_i)) = 0.$$

The variance of Q_2 has bound

$$\begin{aligned}\text{Var}(Q_2) =& \frac{1}{n}\mathbb{E}\left(\left[\int_0^\tau \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)g_0(t;Z_i)dt\right]^2\right)\\ \leq& \frac{1}{n}e^{2K_\theta\tau}\mathbb{E}\left[\int_0^\tau Y_i(t)g_0^2(t;Z_i)dt\right].\end{aligned}$$

Under Assumption 3(b), we have $\text{Var}(Q_2) = O(K_\lambda/n) = o(1)$. By the Tchebychev's inequality, we obtain $Q_2 = o_p(1)$.

Similarly for Q_3, we obtain from Lemma A13 that $\mathbb{E}(Q_3) = 0$. Using the above fact, we give a bound for the variance of Q_3,

$$\begin{aligned}\text{Var}(Q_3) \leq& \frac{1}{n}\mathbb{E}\left(\left[\int_0^\tau \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}dt\right]^2\right)\\ \leq& \frac{1}{n}e^{2K_\theta\tau}\mathbb{E}\left(\int_0^\tau \left[\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}\right]^2 Y_i(t)dt\right).\end{aligned} \tag{A.52}$$

Under Assumption 3(b), we have $\text{Var}(Q_3) = O\left(\left\{\mathcal{M}_\beta\left(\hat{\beta}^{(j)}\right)\right\}^2/n\right) = o(1)$. By the Tchebychev's inequality, we obtain $Q_3 = o_p(1)$.

For Q_4, we also have from Lemma A13

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right).$$

Again using the Cauchy-Schwartz inequality, we bound the total variation of the measure in Q_4,

$$\begin{aligned}&\int_0^\tau [\{\theta_0 + \hat{\beta}^{(j)\top}\check{Z}^{(j)}(t;\hat{\gamma}^{(j)})\}dt + d\widetilde{N}^{(j)}(t;\hat{\gamma}^{(j)})]dt\\ \leq& K_\theta\tau + 1 + \sqrt{\int_0^\tau \frac{n}{\left\{\sum_{i\in I_j} w_i^1(\hat{\gamma}^{(j)})Y_i(t)\right\}^2}dt}\sqrt{e^{2K_\theta\tau}\sum_{i\in I_j} X_i\left(\hat{\beta}^{(j)\top}Z_i\right)^2}.\end{aligned}$$

Using Lemma A14 and the Markov inequality, we have the bound above is of order $O_p\left(\|\hat{\beta}^{(j)}\|_{I_j}\right)$. Therefore, we obtain under Assumption 3b-iii $Q_4 = O_p\left(\|\hat{\beta}^{(j)}\|_{I_j}n^{-\frac{1}{2}}\right) = o_p(1)$.

For terms Q_5, we use the Cauchy-Schwartz inequality

$$|Q_5| \leq \frac{1}{n}\sqrt{\tau\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top \tilde{Z}_i) - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}^2}\sqrt{e^{2K_\theta\tau}\sum_{i\in I_j}\int_0^\tau g_0^2(t;Z_i)Y_i(t)dt}.$$

We apply the Markov's inequality under Assumptions 3(b) and 3b-iii to get

$$Q_5 = O_p\left(\mathcal{D}_{\gamma*}\left(\hat{\gamma}^{(j)}, \gamma_0\right) K_\Lambda\right) = o_p(1).$$

For terms Q_6, we use the Cauchy-Schwartz inequality

$$|Q_6| \leq \sqrt{\frac{\tau}{n}\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top \tilde{Z}_i) - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}^2 \frac{e^{2K_\theta \tau}}{n}\sum_{i\in I_j}\int_0^\tau [\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}]^2 Y_i(t)dt}$$

We apply the Markov's inequality under Assumption 3b-iii to get

$$Q_6 = O_p\left(\mathcal{D}_{\gamma*}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\mathcal{M}_\beta\left(\hat{\beta}^{(j)}\right)\right) = o_p(1).$$

In term Q_7, we establish a uniform bound

$$\begin{aligned}&\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top\tilde{Z}_i) - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right| \\ \leq &\frac{1}{n}\sqrt{\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top\tilde{Z}_i) - \text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i)\}^2}\sqrt{|I_j|e^{2K_\theta\tau}}\end{aligned}$$

by the Cauchy-Schwartz inequality. Hence, the process above is uniformly $O_p\left(\mathcal{D}_{\gamma*}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\right)$. We have the same upper bound for the total variation of the measure as that in Q_4, $O_p\left(\mathcal{M}_\beta\left(\hat{\beta}^{(j)}\right)\right)$. Thus, $Q_7 = O_p\left(\mathcal{D}_{\gamma*}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\mathcal{M}_\beta\left(\hat{\beta}^{(j)}\right)\right) = o_p(1)$.

For terms Q_8 and Q_9, we use the Cauchy-Schwartz inequality to bound the discrepancy between $\mu(t)$ in $\mathcal{M}_\beta\left(\hat{\beta}^{(j)}\right)$ and the empirical $\check{Z}^{(j)}(t, \hat{\beta}^{(j)})$,

$$\begin{aligned}|\hat{\beta}^{(j)\top}\{\mu(t) - \check{Z}^{(j)}(t, \hat{\gamma}^{(j)})\}|^2 &= \left|\sum_{i\in I_j}\frac{w_i^0(\hat{\gamma}^{(j)})Y_i(t)}{\sum_{i'\in I_j} w_{i'}^0(\hat{\gamma}^{(j)})Y_{i'}(t)}\hat{\beta}^{(j)\top}\{\mu(t) - Z_i\}\right|^2 \\ &\leq \frac{\sum_{i'\in I_j}\{w_{i'}^0(\hat{\gamma}^{(j)})Y_{i'}(t)\}^2}{\left\{\sum_{i'\in I_j} w_{i'}^0(\hat{\gamma}^{(j)})Y_{i'}(t)\right\}^2}\sum_{i\in I_j}\left[\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}\right]^2 Y_i(t) \\ &\leq \frac{\sum_{i\in I_j}\left[\hat{\beta}^{(j)\top}\{Z_i - \mu(t)\}\right]^2 Y_i(t)}{\sum_{i'\in I_j} w_{i'}^0(\hat{\gamma}^{(j)})Y_{i'}(t)}. \qquad \text{(A.53)}\end{aligned}$$

The last step above comes from the fact that $w_{i'}^0(\hat{\gamma}^{(j)})Y_{i'}(t) \in [0,1]$. Under Assumption (3b-iii), we obtain

$$\int_0^\tau |\hat{\beta}^{(j)\top}\{\mu(t) - \check{Z}^{(j)}(t, \hat{\gamma}^{(j)})\}|^2 dt = O_p\left(\mathcal{M}_\beta(\hat{\beta}^{(j)})\right) = o_p\left(\mathcal{D}_{\gamma*}(\hat{\gamma}^{(j)})^{-1}\right).$$

Therefore, we follow the strategy of Q_3 and Q_6 to get $Q_8 = o_p(1)$ and $Q_9 = o_p(1)$.

Combining the results for Q_1-Q_9, we establish that $\psi^{(j)}(\theta_0; \hat{\beta}^{(j)}, \hat{\gamma}^{(j)}) = o_p(1)$.

By the Lemma A14, we have the denominator in $\check{\theta}$ (27)

$$Q' = -\frac{1}{n}\sum_{j=1}^{k}\sum_{i\in I_j}(1-D_i)\text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_j)X_i \tag{A.54}$$

is bounded from below by $k\varepsilon_Y/2$.

Along with the identifiability of θ_0 by ψ, we obtain the consistency for $\check{\theta}$.

□

B Auxiliary Results

We state the auxiliary results in Appendices B1-B4, whose proofs are given in Appendix B5. The results in Appendix B1 are technical preliminary steps in the proofs of the main results. We state and prove them separately to promote the conciseness and readability of the proofs of the main results. Appendix B2 contains the classical concentration equalities we use in our proofs. We establish some new concentration results in Appendix B3. We put some minor but frequently used results in Appendix B4. The notations with letter H are all generic and are replaced by suitable objects when we apply the results.

B1 Preliminary Results

Lemma A1. *Under the Assumption 1, we have for θ in a compact neighborhood of θ_0 such that $|\theta| \le K_\theta$*

$$\begin{aligned}
&\sqrt{n}\phi\left(\theta;\hat{\beta},\hat{\Lambda}(\cdot,\theta),\hat{\gamma}\right)\\
=&\sqrt{n}\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0) - \frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0\\
&+o_p(1+\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)+O_p(\sqrt{n}|\theta-\theta_0|^3).
\end{aligned} \tag{A.55}$$

Lemma A2. *Suppose the $|I_j| \asymp n$. Under the Assumption 2, we have for θ in a compact neighborhood of θ_0 such that $|\theta| \le K_\theta$*

$$\begin{aligned}
&\sqrt{n}\phi^{(j)}\left(\theta;\hat{\beta}^{(j)},\hat{\Lambda}^{(j)}(\cdot,\theta),\hat{\gamma}^{(j)}\right)\\
=&\sqrt{n}\phi^{(j)}(\theta_0;\beta_0,\Lambda_0,\gamma_0) - \frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i\in I_j}D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0\\
&+o_p(1+\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)+O_p(\sqrt{n}|\theta-\theta_0|^3).
\end{aligned} \tag{A.55}$$

B2 Classical Concentration Inequalities

Lemma A3. ***Hoeffding's Inequality*** *Theorem 2 p.4 in Hoeffding (1963). If $X_1,\ldots,X_n$ are independent and $a_i \le X_i \le b_i$ $(i=1,2,\ldots,n)$, then for $t>0$*

$$\Pr(\bar{X}-\mu \ge t) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right).$$

Lemma A4. ***A version of Azuma's Inequality*** *Theorem 1 p.3 and Remark 7 p.5 in Sason (2013). Let $\{X_k, \mathcal{F}_k\}_k = 0^\infty$ be a discrete-parameter real-valued martingale sequence such that for every k, the condition $|X_k - X_{k-1}| \leq a_k$ holds almost surely for some non-negative constants $\{a_k\}_{k=1}^\infty$. Then*

$$\Pr\left(\max_{k\in 1,\ldots,n} |X_k - X_0| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n a_k^2}\right)$$

Lemma A5. ***Bernstein Inequality for Sub-exponential Random Variables*** *Chapter 2 Sections 1.3 and 2.2 in Wainwright (2019).*

a) For i.i.d. sample as in Chapter 2 Section 1.3 of Wainwright (2019):

Let X be a random variable with mean $\mathbb{E}(X) = \mu$. If X satisfies the Bernstein's condition with parameter b, i.e.

$$\left|\mathbb{E}\left\{(X-\mu)^k\right\}\right| \leq \frac{1}{2}k!b^k, \text{ for } k = 2, 3, \ldots,$$

the following concentration inequality holds for an i.i.d. sample $X_1, \ldots, X_n$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left\{-\frac{nt^2}{2(b^2+bt)}\right\}.$$

b) For martingale as in Chapter 2 Section 2.2 of Wainwright (2019): Let $M_1, \ldots, M_n$ be a martingale series with respect to filtration $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$. If the martingale differences satisfies the Bernstein's condition with parameter b, i.e.

$$\left|\mathbb{E}\left\{(M_{j+1} - M_j)^k | \mathcal{F}_j\right\}\right| \leq \frac{1}{2}k!b^k, \text{ for } j = 1, \ldots, n-1 \text{ and } k = 2, 3, \ldots,$$

the following concentration inequality holds

$$\mathbb{P}\left(\sup_{j=1,\ldots,n} |M_j| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2(nb^2+bt)}\right\}.$$

Lemma A6. ***Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality*** *(Dvoretzky et al., 1956; Massart, 1990) Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with c.d.f. $F(x)$. Define the empirical c.d.f. as $F_n(x) = n^{-1}\sum_{i=1}^n I(X_i \leq x)$. For any $\varepsilon > 0$,*

$$\Pr\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|\right) \leq 2e^{-2n\varepsilon^2}.$$

B3 New Concentration Results

All the concentration results are adapted to the cross-fitting scheme. We repeated use the following two notations for index set and index set specific filtration.

Definition A1. We denote $I \subset \{1, \ldots, n\}$ be a index set independent of observed data $\{W_i, i = 1, \ldots, n\}$ whose cardinality satisfies $|I| \asymp n$.

Definition A2. We define the filtration for index set I as

$$\mathcal{F}_{I,t} = \sigma\left(\{N_i(u), Y_i(u+), D_i, Z_i : u \leq t, i \in I\} \cup \{\delta_i, X_i, D_i, Z_i : i \in I^c\}\right).$$

Remark A6. The difference between $\mathcal{F}_{I,t}$ with $Y_i(u+)$ and the usual filtration defined with $Y_i(u)$ is that the former contains information about independent out of fold samples and the censoring times at present time t so that the observed censoring times are stopping times with respect to $\mathcal{F}_{I,t}$. On the other hand, we still have the martingale property

$$\mathbb{E}\{M_i(t)|\mathcal{F}_{I,t-}\} = \mathbb{E}\{M_i(t)|\mathcal{F}^*_{I,t-}\} = M_i(t-) \tag{A.56}$$

because the extra censoring information at t is not in $\mathcal{F}_{I,t-}$, and out of fold samples are independent of $M_i(t)$ for $i \in I$.

Lemma A7. *Define the filtration $F_t^{(i)} = \sigma\left(\{N_i(u), Y_i(u), D_i, Z_i : u \le t\}\right)$. Let $H_i(t)$ be a $F_\tau^{(i)}$-measurable random process, satisfying $\mathbb{P}(\sup_{t\in[0,\tau]} |H_i(t)| < K_H) = 1$. Under the model (1) and the Assumption 1-iv,*

$$\mathbb{P}\left(\int_0^\tau H_i(t)Y_i(t)\beta_0^\top Z_i dt > x\right) \tag{A.57}$$

Moreover, we have

$$\left|\int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top Z_i\}dt\right| < 2K_H^2(K_\Lambda + \theta_0 \vee 0)\tau + 4K_H \tag{A.58}$$

and the concentration result for all $\varepsilon \in \left[0, \sqrt{2}\right]$ and index set I defined as in Definition A1

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_i(t)Y_i(t)\beta_0^\top Z_i dt - \int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top Z_i\}dt\right| > K\varepsilon\right) < 4e^{-|I|\varepsilon^2/2}, \tag{A.59}$$

where $K = 2K_H(K_\Lambda + \theta_0 \vee 0)\tau + 2|\mu| + 4K_H$.

Lemma A8. *For an index set I defined as in Definition A1, we define the filtration $\mathcal{F}_{I,t}$ as in Definition A2. Let $M_i(t)$ be the martingale (2) under model (1) and $H_i(t)$ be a nonnegative $\mathcal{F}_{I,t}$-measurable random processes, satisfying $\mathbb{P}(\sup_{t\in[0,\tau]} |H_i(t)| < K_H) = 1$. Denote the order statistics of observed times as $X_{(1)}, \ldots, X_{(|I|)}$. Then,*

$$M_k^H = \frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k)}} H_i(t)dM_i(t),\ k = 0, \ldots, |I| \tag{A.60}$$

is a martingale with respect to $\mathcal{F}_{I,t}$, and we have for $j \ge 2$

$$\left|\mathbb{E}\left\{(M_k^H - M_{k-1}^H)^j|\mathcal{F}_{k-1}^H\right\}\right| \le j!(2K_H/|I|)^j. \tag{A.61}$$

Besides, for every $\varepsilon > K_H/\sqrt{|I|}$ we have

$$\begin{aligned}&|I|\mathbb{E}\left\{(M_k^H - M_{k-1}^H)^2; \sqrt{|I|}|M_k^H - M_{k-1}^H| > \varepsilon\right\}\\ <&(\varepsilon^2|I| + 2K_H\sqrt{|I|} + 2K_H^2)e^{-\varepsilon\sqrt{|I|}/K_H}.\end{aligned} \tag{A.62}$$

Lemma A9. *For an index set I defined as in Definition A1, we define the filtration $\mathcal{F}_{I,t}$ as in Definition A2. Let $M_i(t)$ be the martingale (2) under model (1) and $H_i(t)$ be a $\mathcal{F}_{I,t}$-measurable random processes, satisfying $\mathbb{P}(\sup_{t\in[0,\tau]} |H_i(t)| < K_H) = 1$. Denote $X_{(1)}, \ldots, X_{(|I|)}$ be the order statistics of observed times. Under Assumption 1-iv, for any $\varepsilon < 1$,*

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_i(t)dM_i(t)\right| < 8K_H\varepsilon\right) > 1 - 4e^{-|I|\varepsilon^2/2}. \tag{A.63}$$

Moreover, we also have

$$\bigvee_{t=0}^{\tau}\left\{\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right\} \le \frac{1}{|I|}\sum_{i\in I}\bigvee_{t=0}^{\tau}\int_0^t H_i(u)dM_i(u) < 4K_H + 8K_H\varepsilon \tag{A.64}$$

where $\bigvee_{t=0}^{\tau} f(t)$ *is the total variation of function* $f(t)$ *over* $[0,\tau]$*, and*

$$\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right| < 8K_H\varepsilon + 2K_H/|I| \tag{A.65}$$

whenever the event in (A.63) *occurs.*

Lemma A10. *For an index set* I *defined as in Definition A1, we define the filtration* $\mathcal{F}_{I,t}$ *as in Definition A2. Let* $M_i(t)$ *be the martingale* (2) *under model* (1) *and* $H_i(t)$ *be a* $\mathcal{F}_{I,t}$*-measurable random processes with tight supremum norm* $\max_{i=1,\dots,n}\sup_{t\in[0,\tau]}|H_i(t)| = O_p(1)$*. Under Assumption 1-iv, for any* $\varepsilon < 1$*,*

$$\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{\tau} H_i(t)dM_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right). \tag{A.66}$$

Lemma A11. *Let* H_i *be a random variable, satisfying* $\mathbb{P}\left(\sup_{i=1,\dots,n}|H_i| \le K_H\right) = 1$*. For an index set* I *defined as in Definition A1, we have the concentration result*

$$\mathbb{P}\left(\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}H_iY_i(t) - \mathbb{E}\{H_iY_i(t)\}\right| > 5K_H\varepsilon\right) < 8e^{-|I|\varepsilon^2/2}. \tag{A.67}$$

Lemma A12. *For an index set* I *defined as in Definition A1, we define the filtration* $\mathcal{F}_{I,t}$ *as in Definition A2. Let* $M_i(t)$ *be the martingale* (2) *under model* (1) *and* $H_i(t)$ *be* $\mathcal{F}_{I,t}$*-measurable random processes, satisfying* $\mathbb{P}(\sup_{t\in[0,\tau]}|H_i(t)| < K_H) = 1$*. Let* $\mathcal{H}$ *be a set of functions, potentially not* $\mathcal{F}_{I,t}$*-measurable, but satisfying* $\mathbb{P}\left(\sup_{\tilde{H}\in\mathcal{H}}\sup_{t\in[0,\tau]}|\tilde{H}(t)| < K_V\right) = 1$ *and* $\mathbb{P}\left(\sup_{\tilde{H}\in\mathcal{H}}\bigvee_0^{\tau}\tilde{H}(t) < K_V\right) = 1$*, where* $\bigvee_0^{\tau}$ *is the total variation on* $[0,\tau]$*. Under Assumptions 1-iv and 1-v,*

$$\mathbb{P}\left(\sup_{\tilde{H}\in\mathcal{H}}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{\tau}\tilde{H}(t)H_i(t)dM_i(t)\right| > 16K_HK_V\varepsilon + 2K_HK_V/|I|\right) < 4e^{-|I|\varepsilon^2/2}. \tag{A.68}$$

B4 Other Auxiliary Results

Lemma A13. *Under Assumption* (1-ii) *and models* (1)*, or more general partially linear additive risks model* (8)*, we have*

$$\mathbb{E}[e^{D_i\theta_0 t}Y_i(t)|D_i,Z_i] = \mathbb{E}\{Y_i(t)|Z_i,D_i=0\}. \tag{A.69}$$

Under model (3)*,*

$$\mathbb{E}[\{D_i - \mathrm{expit}(\gamma_0^\top Z_i)\}e^{D_i\theta_0 t}Y_i(t)] = 0 \text{ and } \mathbb{E}[\{D_i - \mathrm{expit}(\gamma_0^\top Z_i)\}e^{D_i\theta_0 t}Y_i(t)Z_i] = \mathbf{0}. \tag{A.70}$$

Moreover, we have for index set I *defined as in Definition A1 under Assumption 1-iii,*

$$\begin{aligned}&\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}\{D_i - \mathrm{expit}(\gamma_0^\top Z_i)\}e^{D_i\theta_0 t}Y_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right) \text{ and}\\ &\sup_{t\in[0,\tau]}\left\|\frac{1}{|I|}\sum_{i\in I}\{D_i - \mathrm{expit}(\gamma_0^\top Z_i)\}e^{D_i\theta_0 t}Y_i(t)Z_i\right\| = O_p\left(\sqrt{\frac{\log(p)}{n}}\right).\end{aligned} \tag{A.71}$$

Lemma A14. *Suppose model (3) is correct, and $\check{\gamma}$ is consistent for γ_0, i.e. $\mathcal{D}_{\gamma*}(\check{\gamma},\gamma_0)=o_p(1)$. For an index set I defined as in Definition A1, we have under Assumption 1-v*

$$\lim_{n\to\infty}\mathbb{P}\left(\inf_{t\in[0,\tau]}\frac{1}{|I|}\sum_{i\in I}w_i^1(\check{\gamma})Y_i(t)>e^{-K_\theta\tau}\varepsilon_Y/2\right)=1 \tag{A.72}$$

$$\textit{and}\ \lim_{n\to\infty}\mathbb{P}\left(\inf_{t\in[0,\tau]}\frac{1}{|I|}\sum_{i\in I}(1-D_i)\mathrm{expit}(\check{\gamma}^\top\tilde{Z}_i)Y_i(t)>\varepsilon_Y/2\right)=1. \tag{A.73}$$

B5 Proofs of the Auxiliary Results

Definition A3. By the mean value theorem for $e^{\theta t}-e^{\theta_0 t}$, we have

$$e^{\theta t}-e^{\theta_0 t}=(\theta-\theta_0)te^{\theta_t t},\ \text{for}\ \theta_t=\xi_t\theta_0+(1-\xi_t)\theta\ \text{with}\ \xi_t\in(0,1). \tag{A.74}$$

In a bounded set of θ such that $|\theta|<K_\theta$, we have the bound $\sup_{t\in[0,\tau]}e^{\theta_t t}\le e^{K_\theta\tau}$. Since θ_t depends on θ, potentially estimated with all information from the data, the process $e^{\theta_t t}$ is not necessarily $\mathcal{F}_{I_j,t}$-adapted, causing extra complication in our proof.

Proof of Lemma A1. We define the filtration as

$$\mathcal{F}_{n,t}=\sigma\left(\{N_i(u),Y_i(u+),D_i,Z_i:u\le t,i=1,\ldots,n\}\right),$$

using $I=\{1,\ldots,n\}$ in Definition A2.

We prove the statement (A.55) by investigating each terms in the following expansion,

$$\begin{aligned}
&\sqrt{n}\phi(\theta;\hat{\beta},\hat{\Lambda}(\cdot,\theta),\hat{\gamma})\\
=\ &\sqrt{n}\phi(\theta;\beta_0,\Lambda_0,\gamma_0)\\
&-n^{-\frac{1}{2}}\sum_{i=1}^n\{D_i-\mathrm{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta t}Y_i(t)(\hat{\beta}-\beta_0)^\top Z_i dt\\
&-n^{-\frac{1}{2}}\sum_{i=1}^n\{D_i-\mathrm{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta t}Y_i(t)\left\{d\hat{\Lambda}(t,\theta)-d\Lambda_0(t)\right\}\\
&-n^{-\frac{1}{2}}\sum_{i=1}^n\{\mathrm{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\mathrm{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta t}dM_i(t;\theta,\beta_0,\Lambda_0)\\
&+n^{-\frac{1}{2}}\sum_{i=1}^n\{\mathrm{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\mathrm{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta t}Y_i(t)(\hat{\beta}-\beta_0)^\top Z_i dt\\
&+n^{-\frac{1}{2}}\sum_{i=1}^n\{\mathrm{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\mathrm{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta t}Y_i(t)\left\{d\hat{\Lambda}(t,\theta)-d\Lambda_0(t)\right\}\\
=\ &Q_1+Q_2+Q_3+Q_4+Q_5+Q_6.
\end{aligned} \tag{A.75}$$

The first term Q_1 contains the leading terms. The rest Q_2-Q_6 are the remainders.

We expand Q_1 with respect to θ at θ_0,

$$\begin{aligned} Q_1 &= \sqrt{n}\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0) \\ &\quad -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{\theta_0 t}D_iY_i(t)dt \\ &\quad +\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{\theta_0 t}D_i t dM_i(t) \\ &\quad +\frac{1}{\sqrt{n}}(\theta-\theta_0)^2\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{\theta_t t}D_i\{t^2dM_i(t)+tY_i(t)dt\} \\ &= Q_{1,1}+Q_{1,2}+Q_{1,3}+Q_{1,4}, \end{aligned} \tag{A.76}$$

where $Q_{1,4}$ comes from the mean value theorem for $e^{\theta t}-e^{\theta_0 t}$ (A.74). $Q_{1,1}$ is the leading term. Each summands in $Q_{1,2}$ is bounded by $e^{\theta_0\tau}$, so $Q_{1,2}$ is of order $O_p(\sqrt{n}|\theta-\theta_0|)$. Through an integral calculation, we have

$$\int_0^\tau e^{D_i\theta_0 t}D_iY_i(t)dt = D_i\int_0^{X_i}e^{\theta_0 t}dt = D_i(e^{\theta_0 X_i}-1)/\theta_0, \tag{A.77}$$

so we can write $Q_{1,2}$ as

$$-\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0. \tag{A.78}$$

In $Q_{1,3}$, we have a $\mathcal{F}_{n,t}$-martingale

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}e^{D_i\theta_0 t}D_i t dM_i(t), \tag{A.79}$$

whose integrand is bounded by $e^{\theta_0\tau}$. By Lemma A9, (A.79) is of order $O_p(n^{-1/2})$. Hence, $Q_{1,3}$ is of order $O_p(|\theta-\theta_0|)=o_p(\sqrt{n}|\theta-\theta_0|)$. Note that we need to prove our statement uniformly in θ, so we cannot directly utilize the martingale structure in $Q_{1,4}$

$$\int_0^\tau e^{\theta_t t}\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}D_i t^2dM_i(t). \tag{A.80}$$

Alternatively, we use Lemma A12 to establish the rate of (A.80) as $O_p(n^{-1/2})$. The other term in $Q_{1,4}$

$$\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{\theta_t t}D_i tY_i(t)dt \tag{A.81}$$

is bounded by $e^{K_\theta\tau}\tau$. Then, $Q_{1,4}$ is of order $O_p(\sqrt{n}|\theta-\theta_0|^2)$. Therefore, we have term Q_1 equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)-\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0 \tag{A.82}$$

plus an $o_p(\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)$ error.

We expand Q_2 with respect to θ,

$$\begin{aligned} Q_2 &= -n^{-\frac{1}{2}} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i \theta_0 t} Y_i(t) (\hat{\beta} - \beta_0)^\top Z_i dt \\ &\quad -n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i=1}^{n} D_i \{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{\theta_t t} Y_i(t) (\hat{\beta} - \beta_0)^\top Z_i dt \\ &= Q_{2,1} + Q_{2,2}, \end{aligned} \tag{A.83}$$

where $Q_{2,2}$ comes from the mean value theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition A3. By the Hölder's inequality, we have an bound for $Q_{2,1}$,

$$|Q_{2,1}| \leq \sqrt{n}\tau \|\hat{\beta} - \beta\|_1 \sup_{t \in [0,\tau]} \left\| \frac{1}{n} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{D_i \theta_0 t} Y_i(t) Z_i \right\|_\infty .$$

From Lemma A13, we have

$$\sup_{t \in [0,\tau]} \left\| \frac{1}{n} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{D_i \theta_0 t} Y_i(t) Z_i \right\|_\infty = O_p\left(\sqrt{\frac{\log(p)}{n}} \right).$$

Under Assumption 1-ix, we have $Q_{2,1} = O_p\left(\sqrt{\log(p)} \|\hat{\beta} - \beta\|_1 \right) = o_p(1)$. We again apply the Hölder's inequality to find the upper bound for $Q_{2,2}$,

$$|Q_{2,2}| \leq \sqrt{n}|\theta - \theta_0|\tau \|\hat{\beta} - \beta\|_1 \sup_{t \in [0,\tau]} \left\| \frac{1}{n} \sum_{i=1}^{n} D_i \{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{\theta_0 t} Y_i(t) Z_i \right\|_\infty .$$

Under Assumptions 1-iii and 1-ix, we have $Q_{2,2} = O_p\left(\sqrt{n}|\theta - \theta_0| \|\hat{\beta} - \beta\|_1 \right) = o\left(\sqrt{n}|\theta - \theta_0| \right)$. Hence, term $Q_2 = Q_{2,1} + Q_{2,2}$ is of order $o_p(\sqrt{n}|\theta - \theta_0| + 1)$.

Very similar to our treatment of Q_2, we expand Q_3 with respect to θ,

$$\begin{aligned} Q_3 &= -\sqrt{n} \int_0^\tau \left[\frac{1}{n} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{D_i \theta_0 t} Y_i(t) \right] \left\{ d\hat{\Lambda}(t, \theta) - d\hat{\Lambda}(t, \theta_0) \right\} \\ &\quad -\sqrt{n} \int_0^\tau \left[\frac{1}{n} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{D_i \theta_0 t} Y_i(t) \right] \left\{ d\hat{\Lambda}(t, \theta_0) - d\Lambda_0(t) \right\} \\ &\quad -n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i=1}^{n} D_i \{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau t e^{\theta_t t} Y_i(t) \left\{ d\hat{\Lambda}(t, \theta) - d\Lambda_0(t) \right\} \\ &= Q_{3,1} + Q_{3,2} + Q_{3,3}, \end{aligned} \tag{A.84}$$

where $Q_{3,3}$ comes from the mean value theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition A3. From Lemma A13, we know that,

$$\sup_{t \in [0,\tau]} \left| \frac{1}{n} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{D_i \theta_0 t} Y_i(t) \right| = O_p\left(n^{-\frac{1}{2}} \right).$$

Together with Assumption 1-viii, the integral $Q_{3,1}$ as an upper bound

$$\sqrt{n}\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right|\bigvee_{t=0}^{\tau}\left\{\hat{\Lambda}(t,\theta)-\hat{\Lambda}(t,\theta_0)\right\}=o_p(\sqrt{n}|\theta-\theta_0|).$$

We apply (10) in Assumption 1-ix to $Q_{3,2}$ and get $Q_{3,2}=o_p(1)$. By Helly-Bray argument (Murphy, 1994), we have a bound for $Q_{3,3}$

$$|Q_{3,3}|\leq\sqrt{n}|\theta-\theta_0|\left\{\left|\hat{\Lambda}(\tau,\theta)-\Lambda_0(\tau)\right|\tau e^{K_\theta\tau}+\int_0^\tau\left|\hat{\Lambda}(t,\theta)-\Lambda_0(t)\right|dte^{\theta_t t}\right\}.$$

Under Assumptions 1-viii and 1-ix, our bound gives the rate $Q_{3,3}=o_p(\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)$. Therefore, $Q_3=Q_{3,1}+Q_{3,2}+Q_{3,3}=o_p(\sqrt{n}|\theta-\theta_0|+1)+O_p(\sqrt{n}|\theta-\theta_0|^2)$.

In terms Q_4-Q_6, we have the model estimation error for the logistic regression. By a mean value theorem argument, we have a uniform bound for the error

$$\left|\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\text{expit}(\gamma_0^\top\tilde{Z}_i)\right|\leq\|\hat{\gamma}-\gamma\|_1\sup_{i=1,\ldots,n}\|Z_i\|_\infty \tag{A.85}$$

because the derivative of function $\text{expit}(\cdot)$ is uniformly bounded by one.

We expand Q_4 with respect to θ,

$$\begin{aligned}Q_4=&-n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\tau\{\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}e^{D_i\theta_0 t}dM_i(t)\\&-n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^{n}\int_0^\tau e^{\theta_t t}D_i\{\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}tdM_i(t)\\&+n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^{n}\{\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau Y_i(t)D_ie^{\theta_t t}(t\theta_t-t\theta_0+1)dt\\=&Q_{4,1}+Q_{4,2}+Q_{4,3},\end{aligned} \tag{A.86}$$

where $Q_{4,3}$ comes from the mean value theorem for $e^{\theta t}-e^{\theta_0 t}$ as in Definition A3. $\hat{\gamma}$ is $\mathcal{F}_{n,t}$-measurable, so we can apply Lemma A10 to $Q_{4,1}$. According to (A.85) and Assumptions 1-iii and 1-ix, $Q_{4,1}=O_p(\|\hat{\gamma}-\gamma\|_1)=o_p(1)$. For $Q_{4,2}$, we apply Lemma A12 with $\mathcal{H}$ be the set of $\{e^{\theta_t t}:|\theta_t|\leq K_\theta\}$ to get $Q_{4,2}=O_p(|\theta-\theta_0|)$. For $Q_{4,3}$, we use the uniform bound from (A.85)

$$|Q_{4,3}|\leq\sqrt{n}|\theta-\theta_0|\|\hat{\gamma}-\gamma\|_1\sup_{i=1,\ldots,n}\|Z_i\|_\infty e^{K_\theta\tau}(2K_\theta\tau+1).$$

Under Assumption 1-iii and 1-ix, $Q_{4,3}=o_p(\sqrt{n}|\theta-\theta_0|)$. Therefore, we obtain $Q_4=o_p(\sqrt{n}|\theta-\theta_0|+1)$.

We apply the Cauchy-Schwartz inequality to Q_5,

$$|Q_5|\leq n^{-\frac{1}{2}}e^{K_\theta\tau}\sqrt{\sum_{i=1}^{n}\{\text{expit}(\hat{\gamma}^\top\tilde{Z}_i)-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}^2}\sqrt{\sum_{i=1}^{n}\left\{(\hat{\beta}-\beta_0)^\top Z_i\right\}^2X_i^2}$$

Hence, we have

$$Q_5=O_p\left(\sqrt{n}\mathcal{D}_\gamma\left(\hat{\gamma},\gamma_0\right)\mathcal{D}_\beta\left(\hat{\beta},\beta_0\right)\right),$$

which is $o_p(1)$ under Assumption 1-ix.

Similarly, we apply the Cauchy-Schwartz inequality to Q_6,

$$|Q_6| \leq n^{-\frac{1}{2}} e^{K_\theta \tau} \sqrt{\sum_{i=1}^{n} \{\text{expit}(\hat{\gamma}^\top \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2} \times \sqrt{\sum_{i=1}^{n} \left[\int_0^\tau e^{D_i \theta t} Y_i(t) \left\{ d\hat{\Lambda}(t,\theta) - d\Lambda_0(t) \right\}\right]^2}.$$

Under Assumption 1-vii, we can apply the Helly-Bray argument (Murphy, 1994) to find the bound,

$$\left|\int_0^\tau e^{D_i \theta t} Y_i(t) \left\{ d\hat{\Lambda}(t,\theta) - d\Lambda_0(t) \right\}\right| \leq \left| e^{D_i \theta X_i} \left\{ \hat{\Lambda}(X_i,\theta) - \Lambda_0(X_i) \right\}\right| + \left| \int_0^{X_i} D_i \theta e^{\theta t} \left\{ \hat{\Lambda}(t,\theta) - \Lambda_0(t) \right\} dt \right|.$$

Hence, $Q_6 = O_p\left(\sqrt{n}\mathcal{D}_\gamma\left(\hat{\gamma},\gamma_0\right) \sup_{t\in[0,\tau]} \left|\hat{\Lambda}(t,\theta) - \Lambda_0\right|\right)$, which is $o_p(1 + \sqrt{n}|\theta - \theta_0|)$ under Assumptions 1-viii and 1-ix.

Combining the results for Q_1-Q_6, we finish the proof. □

Proof of Lemma A2. The proof of the lemma follows fundamentally the same strategy as that of Lemma A1. The main difference is that we use the Cauchy Schwartz inequality instead of the Hölder's inequality to derive MSE type of bounds.

We define the filtration for the j-th fold as

$$\mathcal{F}_{I_j,t} = \sigma\left(\{N_i(u), Y_i(u+), D_i, Z_i : u \leq t, i \in I_j\} \cup \{\delta_i, X_i, D_i, Z_i : i \in I_{-j}\}\right),$$

using $I = I_j$ in Definition A2.

We prove the statement (A.55) by investigating each terms in the following expansion,

$$\begin{aligned}
& \sqrt{n}\phi^{(j)}(\theta; \hat{\beta}^{(j)}, \hat{\Lambda}^{(j)}(\cdot,\theta), \hat{\gamma}^{(j)}) \\
= \; & \sqrt{n}\phi^{(j)}(\theta; \beta_0, \Lambda_0, \gamma_0) \\
& - n^{-\frac{1}{2}} \sum_{i\in I_j} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i\theta t} Y_i(t) (\hat{\beta}^{(j)} - \beta_0)^\top Z_i dt \\
& - n^{-\frac{1}{2}} \sum_{i\in I_j} \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i\theta t} Y_i(t) \left\{ d\hat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t) \right\} \\
& - n^{-\frac{1}{2}} \sum_{i\in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i\theta t} dM_i(t;\theta,\beta_0,\Lambda_0) \\
& + n^{-\frac{1}{2}} \sum_{i\in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i\theta t} Y_i(t) (\hat{\beta}^{(j)} - \beta_0)^\top Z_i dt \\
& + n^{-\frac{1}{2}} \sum_{i\in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i\theta t} Y_i(t) \left\{ d\hat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t) \right\} \\
= \; & Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6. \qquad \text{(A.87)}
\end{aligned}$$

The first term Q_1 contains the leading terms. The rest $Q_2 - Q_6$ are the remainders.

Following exactly the same derivations in the proof of Lemma A1, we have term Q_1 equals

$$\frac{1}{\sqrt{n}}\sum_{i\in I_j}\phi^{(j)}(\theta_0;\beta_0,\Lambda_0,\gamma_0) - \frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i\in I_j}D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}(e^{\theta_0 X_i}-1)/\theta_0 \tag{A.88}$$

plus an $o_p(\sqrt{n}|\theta-\theta_0|) + O_p(\sqrt{n}|\theta-\theta_0|^2)$ error.

We expand Q_2 with respect to θ,

$$\begin{aligned} Q_2 &= -n^{-\frac{1}{2}}\sum_{i\in I_j}\{D_i - \text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt \\ &\quad -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i\in I_j}D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{\theta_t t}Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt \\ &= Q_{2,1}+Q_{2,2}, \end{aligned} \tag{A.89}$$

where $Q_{2,2}$ comes from the mean value theorem for $e^{\theta t}-e^{\theta_0 t}$ as in Definition A3. Denote

$$Q_{2,1,i} = \{D_i - \text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt.$$

Using the independence across folds, we can calculate the expectation for $i\in I_j$

$$\begin{aligned} &\mathbb{E}(Q_{2,1,i}) \\ &= \int_0^\tau \mathbb{E}(\hat{\beta}^{(j)}-\beta_0)^\top\mathbb{E}[\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)Z_i]dt \\ &= \int_0^\tau \mathbb{E}(\hat{\beta}^{(j)}-\beta_0)^\top\mathbb{E}[\mathbb{E}\{D_i-\text{expit}(\gamma_0^\top\tilde{Z}_i)|Z_i\}\mathbb{E}\{e^{D_i\theta_0 t}Y_i(t)|D_i,Z_i\}Z_i]dt, \end{aligned} \tag{A.90}$$

which equals zero by Lemma A13. Hence, $\mathbb{E}(Q_{2,1})=0$. We calculate the variance of $Q_{2,1}$

$$\text{Var}(Q_{2,1}) = n^{-1}\sum_{i\in I_j}\mathbb{E}(Q_{2,1,i}^2) + 2n^{-1}\sum_{i<j,\{i,j\}\subset I_j}\mathbb{E}(Q_{2,1,i}Q_{2,1,j}). \tag{A.91}$$

Note that we have

$$\left|\int_0^\tau e^{D_i\theta t}Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt\right| \le e^{K_\theta\tau}X_i\left|(\hat{\beta}^{(j)}-\beta_0)^\top Z_i\right|. \tag{A.92}$$

Under Assumption 2-i,

$$n^{-1}\sum_{i\in I_j}\mathbb{E}(Q_{2,1,i}^2) \le \frac{|I_j|}{n}e^{2K_\theta\tau}\tau\left\{\mathcal{D}_{\beta*}\left(\hat{\beta}^{(j)},\beta_0\right)\right\}^2 = O_p(r_n^{*2}) = o_p(1).$$

Using the independence across folds again, we have

$$\mathbb{E}(Q_{2,1,i}Q_{2,1,j}) = \mathbb{E}\{\mathbb{E}(Q_{2,1,i}|\hat{\beta}^{(j)})\mathbb{E}(Q_{2,1,j}|\hat{\beta}^{(j)})\} = 0. \tag{A.93}$$

Thus, we establish the rate $\text{Var}(Q_{2,1})=o_p(1)$. By the Tchebychev's inequality, we have $Q_{2,1}=o_p(1)$. For $Q_{2,2}$, we denote

$$Q_{2,2,i} = D_i\{1-\text{expit}(\gamma_0^\top\tilde{Z}_i)\}\int_0^\tau e^{\theta_t t}Y_i(t)(\hat{\beta}^{(j)}-\beta_0)^\top Z_i dt \tag{A.94}$$

apply Cauchy-Schwartz inequality to give an upper bound

$$|Q_{2,2}| \leq n^{-\frac{1}{2}}(\theta - \theta_0)\sqrt{n\sum_{i\in I_j} Q_{2,2,i}^2}. \tag{A.95}$$

Under Assumption 2-i, we have from bound (A.92)

$$\mathbb{E}\{Q_{2,2,i}^2\} \leq e^{2K_\theta \tau}\tau\left\{\mathcal{D}_{\beta*}\left(\hat{\beta}^{(j)}, \beta_0\right)\right\}^2 = o_p(1).$$

Applying the Markov's inequality to $\sum_{i\in I_j} Q_{2,2,i}^2$, we have $Q_{2,2} = o_p(\sqrt{n}|\theta - \theta_0|)$. Hence, term Q_2 is of order $o_p(\sqrt{n}|\theta - \theta_0| + 1)$.

Very similar to our treatment of Q_2, we expand Q_3 with respect to θ,

$$\begin{aligned} Q_3 &= -\sqrt{n}\int_0^\tau \left[\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right]\left\{d\hat{\Lambda}^{(j)}(t,\theta) - d\hat{\Lambda}^{(j)}(t,\theta_0)\right\} \\ &\quad -n^{-\frac{1}{2}}\sum_{i\in I_j}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{d\hat{\Lambda}^{(j)}(t,\theta_0) - d\Lambda_0(t)\right\} \\ &\quad -n^{-\frac{1}{2}}(\theta - \theta_0)\sum_{i\in I_j} D_i\{1 - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}\int_0^\tau te^{D_i\theta_t t}Y_i(t)\left\{d\hat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t)\right\} \\ &= Q_{3,1} + Q_{3,2} + Q_{3,3}, \end{aligned} \tag{A.96}$$

where $Q_{3,3}$ comes from the mean value theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition A3. From Lemma A13, we have,

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right).$$

Together with Assumption 1-viii, the integral $Q_{3,1}$ as an upper bound

$$\sqrt{n}\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right|\bigvee_{t=0}^{\tau}\left\{\hat{\Lambda}^{(j)}(t,\theta) - \hat{\Lambda}^{(j)}(t,\theta_0)\right\} = o_p(\sqrt{n}|\theta - \theta_0|).$$

Denote

$$Q_{3,2,i} = \{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{d\hat{\Lambda}^{(j)}(t,\theta_0) - d\Lambda_0(t)\right\}.$$

Using the independence across folds, we can calculate the expectation for $i \in I_j$ according to Lemma A13

$$\mathbb{E}(Q_{3,2,i}) = \int_0^\tau \mathbb{E}\left(\mathbb{E}[\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)|Z_i]\right)\left[d\mathbb{E}\left\{\hat{\Lambda}^{(j)}(t,\theta_0)\right\} - d\Lambda_0(t)\right],$$

which equals zero by Lemma A13. Hence, $\mathbb{E}(Q_{3,2}) = 0$. Moreover, we have a diminishing bound for $Q_{3,2,i}$ by Helly-Bray argument (Murphy, 1994) under Assumption 2-i

$$\max_{i\in I_j}|Q_{3,2,i}| \leq \left|\hat{\Lambda}^{(j)}(\tau,\theta_0) - \Lambda_0(\tau)\right|e^{K_\theta\tau} + \int_0^\tau \left|\hat{\Lambda}^{(j)}(t,\theta_0) - \Lambda_0(t)\right|de^{\theta_0 t} = o_p(1).$$

We denote $M_{3,2,m} = \frac{1}{\sqrt{n}} \sum_{i \in I_j^{1:m}} Q_{3,2,i}$ as the partial sum of the first m terms in $Q_{3,2}$ whose indices are in $I_j^{1:m}$. It is a martingale with respect to filtration $\mathcal{F}_{3,2,m} = \sigma\left(\{W_i : i \in I_j^{1:m} \cup I_{-j}\}\right)$. By the Azuma's inequality (as in Lemma A4), we have $Q_{3,2} = M_{3,2,|I_j|} = o_p(1)$. Similarly, we apply Helly-Bray argument (Murphy, 1994) to show that

$$|Q_{3,3}| \leq \sqrt{n}|\theta - \theta_0| \left\{ \left| \hat{\Lambda}^{(j)}(\tau, \theta) - \Lambda_0(\tau) \right| \tau e^{K_\theta \tau} + \int_0^\tau \left| \hat{\Lambda}^{(j)}(t, \theta) - \Lambda_0(t) \right| de^{\theta_t t} \right\}.$$

Under Assumptions 1-viii and 2-i, we have $Q_{3,3} = o_p(\sqrt{n}|\theta - \theta_0|) + O_p(\sqrt{n}|\theta - \theta_0|^2)$. Therefore, $Q_3 = Q_{3,1} + Q_{3,2} + Q_{3,3} = o_p(\sqrt{n}|\theta - \theta_0| + 1) + O_p(\sqrt{n}|\theta - \theta_0|^2)$.

We expand Q_4 with respect to θ,

$$\begin{aligned}
Q_4 =& - n^{-\frac{1}{2}} \sum_{i \in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau e^{D_i \theta_0 t} dM_i(t) \\
& - n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i \in I_j} \int_0^\tau e^{\theta_t t} D_i \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} t dM_i(t) \\
& + n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i \in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^\tau Y_i(t) D_i e^{\theta_t t} (t\theta_t - t\theta_0 + 1) dt \\
=& Q_{4,1} + Q_{4,2} + Q_{4,3},
\end{aligned} \tag{A.97}$$

where $Q_{4,3}$ comes from the mean value theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition A3. Denote

$$Q_{4,1,i}(t) = \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} \int_0^t e^{D_i \theta_0 t} dM_i(t).$$

Since $\{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\} e^{D_i \theta_0 t}$ is $\mathcal{F}_{I_j,t}$-adapted, each $Q_{4,1,i}(t)$ is $\mathcal{F}_{I_j,t}$-martingales. Then, $\mathbb{E}\{Q_{4,1}\} = 0$. The optional quadratic variation of $\sum_{i \in I_j} Q_{4,1,i}$ is

$$\begin{aligned}
\left[\sum_{i \in I_j} Q_{4,1,i} \right]_t &= \sum_{i \in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2 \int_0^t e^{2D_i \theta_0 t} dN_i(t) \\
&\leq \sum_{i \in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2 e^{2K_\theta \tau}.
\end{aligned}$$

Under Assumption 2-i, we have $\mathbb{E}\{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2 = \left\{\mathcal{D}_{\gamma *}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\right\}^2 = o_p(1)$. Hence,

$$\text{Var}(Q_{4,1}) = n^{-1} \sum_{i \in I_j} \mathbb{E}\left\{ \left[\sum_{i \in I_j} Q_{4,1,i} \right]_\tau \right\} = o_p(1).$$

We obtain $Q_{4,1} = o_p(1)$ by the Tchebychev's inequality. For $Q_{4,2}$, we apply Lemma A12 with $\mathcal{H}$ be the set of $\{e^{\theta_t t} : |\theta_t| \leq K_\theta\}$ to get $Q_{4,2} = O_p(|\theta - \theta_0|)$. For $Q_{4,3}$, we apply the Cauchy-Schwartz inequality

$$|Q_{4,3}| \leq n^{-\frac{1}{2}} |\theta - \theta_0| \sqrt{\sum_{i \in I_j} \{\text{expit}(\hat{\gamma}^{(j)\top} \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2} \sqrt{n e^{2K_\theta \tau} (K_\theta \tau + \tau)^2}.$$

Again with $\mathbb{E}\{\text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2 = O_p(q_n^*) = o_p(1)$, we obtain from the Markov's inequality that $\sum_{i\in I_j}\{\text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2 = o_p(1)$. Hence, $Q_{4,3} = o_p(\sqrt{n}|\theta - \theta_0|)$. Therefore, we obtain $Q_4 = o_p(\sqrt{n}|\theta - \theta_0| + 1)$.

We apply the Cauchy-Schwartz inequality to Q_5,

$$|Q_5| \le n^{-\frac{1}{2}} e^{K_\theta \tau} \sqrt{\sum_{i\in I_j}\{\text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2} \sqrt{\sum_{i\in I_j}\left\{(\hat{\beta}^{(j)} - \beta_0)^\top Z_i\right\}^2 X_i^2}.$$

Using the independence across folds, we apply the Markov's inequality to get

$$Q_5 = O_p\left(\sqrt{n}\mathcal{D}_{\gamma *}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\mathcal{D}_{\beta *}\left(\hat{\beta}^{(j)}, \beta_0\right)\right),$$

which is $o_p(1)$ under Assumption 2-i.

Similarly, we apply the Cauchy-Schwartz inequality to Q_6,

$$\begin{aligned}|Q_6| \le & n^{-\frac{1}{2}} e^{K_\theta \tau} \sqrt{\sum_{i\in I_j}\{\text{expit}(\hat{\gamma}^{(j)\top}\tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}^2} \\ & \times \sqrt{\sum_{i\in I_j}\left[\int_0^\tau e^{D_i\theta t} Y_i(t)\left\{d\hat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t)\right\}\right]^2}.\end{aligned}$$

Under Assumption 1-vii, we can apply the Helly-Bray argument (Murphy, 1994) to find the bound,

$$\begin{aligned}\left|\int_0^\tau e^{D_i\theta t} Y_i(t)\left\{d\hat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t)\right\}\right| \le & \left|e^{D_i\theta X_i}\left\{\hat{\Lambda}^{(j)}(X_i,\theta) - \Lambda_0(X_i)\right\}\right| \\ & + \left|\int_0^{X_i} D_i\theta e^{\theta t}\left\{\hat{\Lambda}^{(j)}(t,\theta) - \Lambda_0(t)\right\} dt\right|.\end{aligned}$$

Hence, $Q_6 = O_p\left(\sqrt{n}\mathcal{D}_{\gamma *}\left(\hat{\gamma}^{(j)}, \gamma_0\right)\sup_{t\in[0,\tau]}\left|\hat{\Lambda}^{(j)}(t,\theta) - \Lambda_0\right|\right)$, which is $o_p(1 + \sqrt{n}|\theta - \theta_0|)$ under Assumptions 1-viii and 2-i.

Combining the results for Q_1-Q_6, we finish the prove. □

Proof of Lemma A.7. We prove the result for nonnegative $H_i(t)$. The general result can be obtained through decomposing $H_i(t)$ into the difference of two nonnegative processes

$$H_i(t) = H_i(t) \vee 0 - [-\{H_i(t) \wedge 0\}] \tag{A.98}$$

and use the union bound with the result for the nonnegative processes.

Under the model (1), μ satisfies $\mathbb{P}(D_i\theta_0 + \beta_0^\top Z_i \ge -d\Lambda_0(t)) = 1$. By the Assumption 1-iv, we have a lower bound $\mathbb{P}(\beta_0^\top Z_i > -K_\Lambda - \theta_0 \vee 0) = 1$. The $\beta_0^\top Z_i$ is potentially unbounded from above, so we have to study the bound for the upper tail. For $x > K_H(K_\Lambda + \theta_0 \vee 0)\tau$,

$$\begin{aligned}& \mathbb{P}\left(\int_0^\tau H_i(t)Y_i(t)\beta_0^\top Z_i dt > x\right) \\ \le\ & \mathbb{P}\left(K_H X_i \beta_0^\top Z_i > x\right) \\ \le\ & \mathbb{E}\left[I(\beta_0^\top Z_i > K_\Lambda + \theta_0 \vee 0) I(C_i > x/K_H)\exp\left\{-\frac{x}{K_H}\frac{D_i\theta_0 + \beta_0^\top Z_i}{\beta_0^\top Z_i} - \Lambda_0\left(\frac{x/K_H}{\beta_0^\top Z_i}\right)\right\}\right] \\ \le\ & e^{-x/(2K_H)}.\end{aligned} \tag{A.99}$$

Denote $A_i = \int_0^\tau H_i(t)Y_i(t)\beta_0^\top Z_i dt$, $\mu = \int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top Z_i\}dt$ and $K_A = K_H(K_\Lambda + \theta_0 \vee 0)\tau$. First, we can find a bound for the expectation

$$\begin{aligned}
|\mu| &= \left|\int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top Z_i\}dt\right| \\
&\leq \left|\int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top Z_i I(|\beta_0^\top Z_i| < K_A)\}dt\right| + \left|\mathbb{E}\left\{\int_0^\tau H_i(t)Y_i(t)\beta_0^\top Z_i I(\beta_0^\top Z_i \geq K_A)dt\right\}\right| \\
&\leq K_H K_A \tau + \int_0^\infty \mathbb{P}\left(\int_0^\tau H_i(t)Y_i(t)\beta_0^\top Z_i I(\beta_0^\top Z_i \geq K_A)dt > x\right)dx \\
&\leq K_H K_A + 2K_H. \qquad \text{(A.100)}
\end{aligned}$$

Then, we bound the centered moments for $k \geq 2$

$$\begin{aligned}
\mathbb{E}(A_i - \mu)^k &= \mathbb{E}\{(A_i - \mu)^k I(A_i < K_A + \mu \vee 0)\} + \mathbb{E}\{(A_i - \mu)^k I(A_i \geq K_A + \mu \vee 0)\} \\
&\leq (K_A + |\mu|)^k + \int_0^\infty \mathbb{P}\{(A_i - \mu)^k I(A_i \geq K_A + \mu \vee 0) > x\}dx \\
&\leq (K_A + |\mu|)^k + \int_0^{(K_A - \mu \wedge 0)^k} \mathbb{P}(A_i \geq K_A + \mu \vee 0)dx \\
&\quad + \int_{(K_A - \mu \wedge 0)^k}^\infty \mathbb{P}(A_i > x^{1/k} + \mu)dx \\
&\leq 2(K_A + |\mu|)^k + k!(2K_H)^k \\
&\leq k!(K_A + |\mu| + 2K_H)^k \qquad \text{(A.101)}
\end{aligned}$$

Thus, A_i is sub-exponential. By Bernstein inequality for sub-exponential random variables (as in Lemma A5), we have for any $\varepsilon \in [0, \sqrt{2}]$

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I} A_i - \mu\right| > \varepsilon(K_A + |\mu| + 2K_H)\right) < 2e^{-|I|\varepsilon^2/2}. \qquad \text{(A.102)}$$

We thus complete the proof. □

Proof of Lemma A8. Let $X_{(1)}, \ldots, X_{(|I|)}$ be the order statistics of observed times. Under filtration $\mathcal{F}_{I,t}$, they are ordered stopping times (see Definition A2 and Remark A6). By optional stopping theorem (Durrett, 2013), we construct a discrete stopped martingale

$$M_k^H = \frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k)}} H_i(t)dM_i(t) \qquad \text{(A.103)}$$

under filtration $\mathcal{F}_k^H = \sigma\{N_i(u), Y_i(u+), D_i, Z_i, X_{(k)} : u \in [0, X_{(k)}], i \in I\}$. The increment of the discrete martingale has two components,

$$\begin{aligned}
M_k^H - M_{k-1}^H &= \frac{1}{|I|}\sum_{i\in I} H_i(X_{(k)})dN_i(X_{(k)}) \\
&\quad - \frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)})\int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top Z_i\}dt + d\Lambda_0(t)], \qquad \text{(A.104)}
\end{aligned}$$

one from the jumps of $N_i(t)$ and the other from the compensator. Under Assumption 1-iv, there is almost surely no ties in the observed event times, so we have a bound

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I} H_i(X_{(k)})dN_i(X_{(k)})\right| \le K_H/|I|\right) = \mathbb{P}\left(\frac{1}{|I|}\max_{i\in I} H_i(X_{(k)}) \le K_H/|I|\right) = 1. \tag{A.105}$$

The compensator term in (A.104), second on the right hand side, is potentially unbounded. We have to study its tail distribution. Conditioning on $\mathcal{F}^H_{k-1}$, we calculate the distribution of $X_{(k)}$ as

$$\begin{aligned}
& \mathbb{P}(X_{(k)} \ge X_{(k-1)} + x|\mathcal{F}^H_{k-1}) \\
= & \prod_{i=1}^{|I|} \mathbb{P}(C_i \wedge T_i \ge X_{(k-1)} + x | C_i \wedge T_i \ge X_{(k-1)})^{Y_i(X_{(k-1)})} \\
\le & \exp\left[-\sum_{i\in I} Y_i(X_{(k-1)})\{(D_i\theta_0 + \beta_0^\top Z_i)x + \Lambda_0(X_{(k-1)}+x) - \Lambda_0(X_{(k-1)})\}\right].
\end{aligned} \tag{A.106}$$

We denote the function in the exponential index as

$$A(x) = \sum_{i\in I} Y_i(X_{(k-1)})\{(D_i\theta_0 + \beta_0^\top Z_i)x + \Lambda_0(X_{(k-1)}+x) - \Lambda_0(X_{(k-1)})\}. \tag{A.107}$$

Note that $A(x)$ is nondecreasing, so its inverse $A^{-1}(x)$ is well defined. Next, we evaluate the tail distribution of the compensator term

$$\begin{aligned}
& \mathbb{P}\left(\frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top Z_i\}dt + d\Lambda_0(t)] \ge x \middle| \mathcal{F}^H_{k-1}\right) \\
\le & \mathbb{P}(K_H A(X_{(k)} - X_{(k-1)})/|I| \ge x) \\
= & \mathbb{P}\{X_{(k)} \ge X_{(k-1)} + A^{-1}(nx/K_H)\} \\
\le & e^{-|I|x/K_H}.
\end{aligned} \tag{A.108}$$

For $j \ge 2$, we calculate the moments

$$\begin{aligned}
& \left|\mathbb{E}\left\{(M^H_k - M^H_{k-1})^j | \mathcal{F}^H_{k-1}\right\}\right| \\
\le & \left[\mathbb{E}\left\{\left|\frac{1}{|I|}\sum_{i\in I} H_i(X_{(k)})dN_i(X_{(k)})\right|^j \middle| \mathcal{F}^H_{k-1}\right\}^{\frac{1}{j}}\right. \\
& \left. + \mathbb{E}\left\{\left|\frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top Z_i\}dt + d\Lambda_0(t)]\right|^j \middle| \mathcal{F}^H_{k-1}\right\}^{\frac{1}{j}}\right]^j \\
\le & \left[\frac{K_H}{|I|} + \left\{\int_0^\infty e^{-|I|x^{\frac{1}{j}}/K_H} dx\right\}^{\frac{1}{j}}\right]^j \\
= & \left[\frac{K_H}{|I|} + \frac{K_H}{|I|}(j!)^{\frac{1}{j}}\right]^j \\
\le & j!(2K_H/|I|)^j.
\end{aligned} \tag{A.109}$$

This statement above proves (A.61), the first conclusion of the lemma.

For $\varepsilon > K_H/\sqrt{|I|}$, event

$$\sqrt{|I|}|M_k^H - M_{k-1}^H| > \varepsilon \tag{A.110}$$

occurs only if the following event occurs,

$$\begin{aligned} & \frac{1}{|I|}\sum_{i\in I} H_i(X_{(k)})dN_i(X_{(k)}) + \varepsilon/\sqrt{|I|} \\ < \; & \frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top Z_i\}dt + d\Lambda_0(t)]. \end{aligned} \tag{A.111}$$

We can bound

$$\begin{aligned} & \mathbb{E}\left\{(M_k^H - M_{k-1}^H)^2; \sqrt{|I|}|M_k^H - M_{k-1}^H| > \varepsilon\right\} \\ \leq \; & \mathbb{E}\Bigg\{ \left(\frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top Z_i\}dt + d\Lambda_0(t)]\right)^2 \\ & \quad \times I\left(\frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top Z_i\}dt + d\Lambda_0(t)] > \varepsilon/\sqrt{|I|}\right)\Bigg\} \\ \leq \; & \frac{\varepsilon^2}{|I|}e^{-\varepsilon\sqrt{|I|}/K_H} + \int_{\varepsilon^2/|I|}^{\infty} e^{-|I|\sqrt{x}/K_H}dx \\ = \; & \frac{\varepsilon^2|I| + 2K_H\sqrt{|I|} + 2K_H^2}{|I|^2}e^{-\varepsilon\sqrt{|I|}/K_H}. \end{aligned} \tag{A.112}$$

This proves (A.62), the other conclusion of the lemma. □

Proof of Lemma A9. Without loss of generality, we again prove the result for the nonnegative $H_i(t)$.

Let $X_{(1)}, \ldots, X_{(|I|)}$ be the order statistics of observed times. We define the sequence M_k^H, $k = 1, \ldots, n$, along with filtration $\mathcal{F}_k^* = \mathcal{F}_{I,X_{(k)}}$, as in Lemma A8. By Lemma A8, M_k^H is a $\mathcal{F}_k^*$-martingale satisfying (A.61), so we can apply the Bernstein's inequality for martingale differences (as in Lemma A5). For $\varepsilon < 1$, we have

$$\mathbb{P}\left(\sup_{k=1,\ldots,|I|}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(i)}} H_i(t)dM_i(t)\right| > 4K_H\varepsilon\right) = \mathbb{P}\left(\sup_{k=1,\ldots,|I|}|M_k^H| > 4K_H\varepsilon\right) < 2e^{-|I|\varepsilon^2/2}. \tag{A.113}$$

This proves (A.63), the first result of the lemma.

The total variation of the integral with nonnegative $H_i(t)$'s can be written as

$$\begin{aligned} \bigvee_{t=0}^{\tau}\left\{\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right\} &= \frac{1}{|I|}\sum_{i\in I}\bigvee_{t=0}^{\tau}\int_0^t H_i(u)dM_i(u) \\ &= \frac{2}{|I|}\sum_{i\in I}\int_0^{\tau} H_i(u)dN_i(u) - \frac{1}{|I|}\sum_{i\in I}\int_0^{\tau} H_i(u)dM_i(u). \end{aligned} \tag{A.114}$$

Hence, (A.64) the second result of the lemma follows directly from the first result (A.113).

To find the bound of variation between $X_{(k-1)}$ and $X_{(k)}$, simply consider that $H_i(t)$ is nonnegative while $dN_i(t)$ and $Y_i(t)\{(D_i\theta_0 + \beta_0^\top Z_i)dt + d\Lambda_0(t)\}$ are nonnegative measures. Hence, the extremal values in the intervals can be explicitly expressed as

$$\sup_{t\in[X_{(k-1)},X_{(k)})} \frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u) = \frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k-1)}} H_i(u)dM_i(u) = M_{k-1}^H, \tag{A.115}$$

and

$$\inf_{t\in[X_{(k-1)},X_{(k)})} \frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u) = \frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k)}-} H_i(u)dM_i(u) = M_k^H - \frac{H_{i_k}(X_{(k)})}{|I|}. \tag{A.116}$$

Therefore,

$$\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right| \le \sup_{k=1,\ldots,n}|M_k^H| + K_H/|I|. \tag{A.117}$$

For general $H_i(t)$, we simply decompose $H_i(t) = H_i^+(t) - H_i^-(t)$ and use the union bound. □

Proof of Lemma A10. The proof uses the conclusion of Lemma A9. For any $\varepsilon > 0$, we can find K_ε according to the tightness of $H_i(t)$ such that $\mathbb{P}\left(\max_{i=1,\ldots,n}\sup_{t\in[0,\tau]}|H_i(t)| > K_\varepsilon\right) < \varepsilon/2$. Define the truncated processes $H_{i,\varepsilon}(t) = (-K_\varepsilon)\vee\{H_i(t)\wedge K_\varepsilon\}$, which is still $\mathcal{F}_{I,t}$-adapted, as well as bounded by K_ε. By Lemma A9, we have

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_{i,\varepsilon}(t)dM_i(t)\right| < 8K_\varepsilon\frac{\log(8/\varepsilon)}{\sqrt{|I|/2}}\right) > 1-\varepsilon/2.$$

Since $H_{i,\varepsilon}(t) = H_i(t)$ for all $i = 1,\ldots,n$ and $t\in[0,\tau]$ with probability at least $1-\varepsilon/2$, we have

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_i(t)dM_i(t)\right| < 8K_\varepsilon\frac{\log(8/\varepsilon)}{\sqrt{|I|/2}}\right) > 1-\varepsilon.$$

The last equation defines the rate in (A.66). □

Proof of Lemma (A11). Let B_i, $i\in I$, be independent Bernoulli random variables with rate $(H_i + K_H)/(2K_H)$. By a simple calculation, we have the following empirical distribution for B_iX_i

$$\frac{1}{|I|}\sum_{i\in I}B_iY_i(t) = \frac{1}{|I|}\sum_{i\in I}I(B_iX_i\ge t) \text{ and } \mathbb{E}\{B_iY_i(t)\} = \frac{1}{2K_H}\mathbb{E}\{H_iY_i(t)\} + \frac{1}{2}\mathbb{E}\{Y(t)\}. \tag{A.118}$$

We decompose

$$\begin{aligned}\frac{1}{|I|}\sum_{i\in I}H_iY_i(t) - \mathbb{E}\{H_iY_i(t)\} &= \frac{2K_H}{|I|}\sum_{i\in I}B_iY_i(t) - \mathbb{E}\{H_iY_i(t)\} - K_H\mathbb{E}\{Y(t)\}\\ &\quad -\frac{K_H}{|I|}\sum_{i\in I}Y_i(t) + K_H\mathbb{E}\{Y(t)\}\\ &\quad -\frac{2K_H}{|I|}\sum_{i\in I}\left(B_i - \frac{H_i+K_H}{2K_H}\right)Y_i(t).\end{aligned} \tag{A.119}$$

Applying the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (as in Lemma A6) to the first two terms in (A.119), we have

$$\mathbb{P}\left(\sup_{t\in[0,\tau]}\left|\frac{2K_H}{|I|}\sum_{i\in I}B_iY_i(t)-\mathbb{E}\{H_iY_i(t)\}-K_H\mathbb{E}\{Y(t)\}\right|>K_H\varepsilon\right)\leq 2e^{-|I|\varepsilon^2/2} \tag{A.120}$$

$$\text{and } \mathbb{P}\left(\sup_{t\in[0,\tau]}\left|\frac{K_H}{|I|}\sum_{i\in I}Y_i(t)-K_H\mathbb{E}\{Y(t)\}\right|>K_H\varepsilon\right)\leq 2e^{-|I|\varepsilon^2/2}. \tag{A.121}$$

Denote $X_{(i)}$, $i=1,\ldots,n$, as the order statistics of X_i's. We further decompose the third term in (A.119) as

$$\begin{aligned}\frac{2K_H}{|I|}\sum_{i\in I}\left(B_i-\frac{H_i+K_H}{2K_H}\right)Y_i(X_{(k)}) &= \frac{2K_H}{|I|}\sum_{i\in I}\left(B_i-\frac{H_i+K_H}{2K_H}\right)\\ &\quad -\frac{2K_H}{|I|}\sum_{i=1}^{k}\left(B_{(i)}-\frac{H_{(i)}+K_H}{2K_H}\right).\end{aligned} \tag{A.122}$$

By the Hoeffding's inequality (as in Lemma A3), we bound the first term in (A.122)

$$\mathbb{P}\left(\left|\frac{2K_H}{|I|}\sum_{i\in I}\left(B_i-\frac{H_i-K_H}{2K_H}\right)\right|>K_H\varepsilon\right)<2e^{-|I|\varepsilon^2/2}. \tag{A.123}$$

Let (i) be the i-th element in fold I. We define a filtration $\mathcal{F}_m^H=\sigma(\{(H_i,X_i):i\in I\}\cup\{B_{(i)}:i=1,\ldots,m\})$ under which we have the following martingale

$$M_m^H=\frac{2K_H}{|I|}\sum_{i=1}^{m}\left(B_{(i)}-\frac{H_{(i)}+K_H}{2K_H}\right). \tag{A.124}$$

By the Azuma's inequality (as in Lemma A4), we have

$$\mathbb{P}\left(\left|M_{|I|}^H\right|>2K_H\varepsilon\right)<2e^{-|I|\varepsilon^2/2}. \tag{A.125}$$

We finish the proof by putting the concentration inequalities together. □

Proof of Lemma A12. By Lemma A9, the probability that the event

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau H_i(u)dM_i(u)<8K_H\varepsilon \tag{A.126}$$

is no less than $1-4e^{-n\varepsilon^2/2}$. We shall show that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\tilde{H}(t)H_i(t)dM_i(t)\right|<16K_HK_V\varepsilon+2K_HK_V/n \tag{A.127}$$

on such event.

By Lemma A9, the following function

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(u)dM_i(u) \tag{A.128}$$

has total variation bounded by $4K_H + 8K_H\varepsilon$ on event (A.127). As a result, we can apply the Helly-Bray integration by parts (Murphy, 1994)

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \tilde{H}(t)H_i(t)dM_i(t) = \frac{\tilde{H}(\tau)}{n}\sum_{i=1}^{n}\int_0^\tau H_i(t)dM_i(t) - \int_0^\tau \left\{\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(u)dM_i(u)\right\} d\tilde{H}(t). \tag{A.129}$$

By Lemma A9, both terms have bound on event (A.127)

$$\left|\frac{\tilde{H}(\tau)}{n}\sum_{i=1}^{n}\int_0^\tau H_i(t)dM_i(t)\right| \le K_V \times 8K_H\varepsilon, \tag{A.130}$$

$$\left|\int_0^\tau \left\{\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(u)dM_i(u)\right\} d\tilde{H}(t)\right| \le K_V \times (8K_H\varepsilon + 2K_H/n). \tag{A.131}$$

Plugging in the upper bounds to (A.129) finish the proof. □

Proof of Lemma A13. Since we assume that T_i and C_i are independent given D_i and Z_i, we have

$$\mathbb{E}[Y_i(t)|D_i, Z_i] = \mathbb{P}(T_i \wedge C_i \ge t|D_i, Z_i) = \mathbb{P}(C_i \ge t|D_i, Z_i)\mathbb{P}(T_i \ge t|D_i, Z_i). \tag{A.132}$$

Under Assumption 1-ii, the censoring time is independent of treatment given covariates, so

$$\mathbb{P}(C_i \ge t|D_i, Z_i) = \mathbb{P}(C_i \ge t|Z_i) \tag{A.133}$$

is $\sigma\{Z_i\}-$measurable. Under model (8),

$$\mathbb{P}(T_i \ge t|D_i, Z_i) = e^{\int_0^t \lambda(t;D_i,Z_i)dt} = e^{-D_i\theta_0 t - \int_0^t g_0(t;Z_i)dt} = e^{-D_i\theta_0 t}\mathbb{P}(T_i \ge t|D_i = 0, Z_i). \tag{A.134}$$

Therefore, we have the following representation

$$\mathbb{E}[e^{D_i\theta_0 t}Y_i(t)|D_i, Z_i] = \mathbb{P}(C_i \ge t|Z_i)e^{-\int_0^t g_0(t;Z_i)dt} = \mathbb{E}\{Y_i(t)|Z_i, D_i = 0\}, \tag{A.135}$$

which is obviously $\sigma\{Z_i\}-$measurable. By the tower property of conditional expectation, we can calculate the expectations for any $\sigma\{Z_i\}$-measurable random variable U_i through

$$\begin{aligned} & \mathbb{E}[\{D_i - \text{expit}(\gamma_0^\top Z_i)\}e^{D_i\theta_0 t}Y_i(t)U_i] \\ = \ & \mathbb{E}[\{D_i - \text{expit}(\gamma_0^\top Z_i)\}\mathbb{E}\{e^{D_i\theta_0 t}Y_i(t)|D_i, Z_i\}U_i] \\ = \ & \mathbb{E}[\mathbb{E}\{D_i - \text{expit}(\gamma_0^\top Z_i)|Z_i\}\mathbb{E}\{Y_i(t)|Z_i, D_i = 0\}U_i] \\ = \ & 0. \end{aligned} \tag{A.136}$$

We obtain the two equations in (A.70) by setting U_i above as 1 and Z_i, respectively.

To deliver the concentration result (A.71), we decompose

$$\frac{1}{|I|}\sum_{i\in I}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t) = \frac{1}{|I|}\sum_{i\in I} D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t) - \frac{1}{|I|}\sum_{i\in I}(1-D_i)\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t).$$

Each coordinate of

$$\frac{1}{|I|}\sum_{i\in I} D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t) \text{ and } \frac{1}{|I|}\sum_{i\in I}\text{expit}(\gamma_0^\top \tilde{Z}_i)Y_i(t),$$

is bounded, so we can apply Lemma A11 to get

$$\sup_{t\in[0,\tau]}\left|e^{\theta_0 t}\frac{1}{|I|}\sum_{i\in I} D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t) - e^{\theta_0 t}\mathbb{E}\left[D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t)\right]\right| = O_p\left(n^{-\frac{1}{2}}\right),$$
$$\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}(1-D_i)\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t) - \mathbb{E}\left[(1-D_i)\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t)\right]\right| = O_p\left(n^{-\frac{1}{2}}\right). \quad \text{(A.137)}$$

From (A.70), we know that

$$e^{\theta_0 t}\mathbb{E}\left[D_i\{1-\text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(t)\right] = \mathbb{E}\left[(1-D_i)\text{expit}(\gamma_0^\top \tilde{Z}_i)Y_i(t)\right]. \quad \text{(A.138)}$$

Therefore, we have proved the first rate in (A.71) by combining (A.137) and (A.138). In the same way under Assumption 1-iii, we have a concentration result from Lemma A11 for each coordinate of $\frac{1}{|I|}\sum_{i\in I}\{D_i - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)Z_i$. We take the union bound to obtain the second rate in (A.71). □

Proof of Lemma A14. We provide the proof for the first result (A.72). The proof for the second result (A.73) is identical. Since the weights $w_i^1(\check{\gamma})$ are nonnegative and $Y_i(t)$'s are non-increasing, we have lower bound

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\check{\gamma})Y_i(t) \geq \frac{1}{|I|}\sum_{i\in I} D_i\{1-\text{expit}(\check{\gamma}^\top \tilde{Z}_i)\}Y_i(\tau). \quad \text{(A.139)}$$

it is sufficient to show

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{1}{|I|}\sum_{i\in I} w_i^1(\check{\gamma})Y_i(\tau) > \varepsilon_Y/2\right) = 1. \quad \text{(A.140)}$$

We decompose

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\check{\gamma})Y_i(\tau) = \frac{1}{|I|}\sum_{i\in I} w_i^1(\gamma_0)Y_i(\tau) - \frac{1}{|I|}\sum_{i\in I} D_i\{\text{expit}(\check{\gamma}^\top \tilde{Z}_i) - \text{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(\tau). \quad \text{(A.141)}$$

The first term in (A.141) has expectation bounded away from zero by Assumption 1-v

$$\mathbb{E}\{w_i^1(\gamma_0)Y_i(\tau)\} = \mathbb{E}\{\text{Var}(D_i|\tilde{Z}_i)e^{\theta_0 t}\mathbb{E}\{Y_i(\tau)|\tilde{Z}_i, D_i=0\}\} \geq e^{-K_\theta \tau}\varepsilon_Y. \quad \text{(A.142)}$$

Since $w_i^1(\gamma_0)Y_i(\tau)$ are i.i.d. random variables in $[0,1]$, we have by Hoeffding's inequality (as in Lemma A3),

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\gamma_0)Y_i(\tau) = \mathbb{E}\{\mathrm{Var}(D_i|\tilde{Z}_i)e^{\theta_0 t}\mathbb{E}\{Y_i(\tau)|\tilde{Z}_i, D_i=0\}\} + O_p(n^{-1/2}) \geq e^{-K_\theta \tau}\varepsilon_Y + o_p(1). \quad \text{(A.143)}$$

By the Cauchy-Schwartz inequality, we have the bound for the second term in (A.141),

$$\begin{aligned} &\left|\frac{1}{|I|}\sum_{i\in I} D_i\{\mathrm{expit}(\check{\gamma}^\top \tilde{Z}_i) - \mathrm{expit}(\gamma_0^\top \tilde{Z}_i)\}Y_i(\tau)\right| \\ \leq\ &\sqrt{\frac{1}{|I|}\sum_{i\in I}\{\mathrm{expit}(\check{\gamma}^\top \tilde{Z}_i) - \mathrm{expit}(\gamma_0^\top \tilde{Z}_i)\}^2}. \end{aligned} \quad \text{(A.144)}$$

By the Markov's inequality, the bound above is of order $O_p\left(\mathcal{D}_{\gamma*}(\check{\gamma}, \gamma_0)\right) = o_p(1)$. Therefore, we have

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\check{\gamma})Y_i(\tau) + o_p(1) \geq \varepsilon_Y. \quad \text{(A.145)}$$

Hence, we obtain (A.140), a sufficient condition for (A.72). □

References

P. K. Andersen and R. Gill. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10(4):1100–1120, 1982.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society, Serie B*, 80(4):597–623, 2018.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650, 2013.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102:77–94, 2015.

P. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, 1998.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

J. Bradic, J. Fan, and J. Jiang. Regularization for Coxs proportional hazards model with NP-dimensionality. *Annals of Statistics*, 39(6):3092–3120, 2011.

J. Bradic, J. Fan, and Y. Zhu. Testability of high-dimensional linear models with non-sparse structures. *arXiv*, 2018.

J. Bradic, S. Wager, and Y. Zhu. Sparsity double robust inference of average treatment effects. *arXiv*, 2019.

B. Brown. Martingale central limit theorems. *The Annals of Mathematical Statistics*, 42(1):59–66, 1971.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21 (1):C1–C68, 2018a.

V. Chernozhukov, W. Newey, and J. Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv*, 2018b.

D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

R. Dezeure, P. Bhlmann, L. Meier, and N. Meinshausen. High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30(4):533–558, 2015.

O. Dukes, T. Martinussen, E. J. Tchetgen Tchetgen, and S. Vansteelandt. On doubly robust estimation of the hazard difference. *Biometrics*, 75:100–019, 2019.

R. Durrett. *Probability: Theory and Examples, 4th edition.* Cambridge University Press, 2013.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27(3):642–669, 1956.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189:1–23, 2015.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv*, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010.

S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

A. Gorst-Rasmussen and T. Scheike. Coordinate descent methods for the penalized semiparametric additive hazards model. *Journal of Statistical Software, Articles*, 47(9):1–17, 2012.

J. Hadley, K. R. Yabroff, M. J. Barrett, D. F. Penson, C. S. Saigal, and A. L. Potosky. Comparative effectiveness of prostate cancer treatments: Evaluating statistical adjustments for confounding in observational data. *Journal of the National Cancer Institute*, 103:1780–1793, 2010.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):2546, 2012.

L. C. Harlan, A. Potosky, F. D. Gilliland, R. Hoffman, P. C. Albertsen, A. S. Hamilton, J. W. Eley, J. L. Stanford, and R. A. Stephenson. Factors associated with initial therapy for clinically localized prostate cancer: Prostate cancer outcomes study. *Journal of the National Cancer Institute*, 93(24):1864–1871, 2001.

N. L. Hjort. Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics*, 13(2):63–85, 1986.

N. L. Hjort. On inference in parametric survival data models. *International Statistical Review*, 60(3): 355–387, 1992.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. Hou, A. Paravati, J. Hou, R. Xu, and J. Murphy. High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data. *Statistics in Medicine*, 37(4): 3486–3502, 2018.

J. Hou, J. Bradic, and R. Xu. Inference under fine-gray competing risks model with high-dimensional covariates. *Electronic Journal of Statistics*, page to appear, 2019.

J. Huang, T. Sun, Z. Ying, Y. Yu, and C.-H. Zhang. Oracle inequalities for the LASSO in the Cox model. *Annals of Statistics*, 41(3):1142–1165, 2013.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society, Serie B*, 76:243–263, 2014.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

R. Jiang, W. Lu, R. Song, M. G. Hudgens, and S. Naprvavnik. Doubly robust estimation of optimal treatment regimes for survival datawith application to an hiv/aids study. *The Annals of Applied Statistics*, 11(3):1763–1786, 2017.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data (2nd ed.)*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

S. Kang, W. Lu, and J. Zhang. On estimation of the optimal treatment regime with the additive hazards model. *Statistica Sinica*, 28(3):1539–1560, 2018.

C. Leng and S. Ma. Path consistent model selection in additive risk model via Lasso. *Statistics in Medicine*, 26:3753–3770, 2007.

D. Y. Lin and Z. Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994.

W. Lin and J. Lv. High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264, 2013.

G. Lu-Yao, T. A. Stukel, and S.-L. Yao. Changing patterns in competing causes of death in men with prostate cancer: a population based study. *The Journal of Urology*, 171(6):2285–2290, 2004.

T. Martinussen and T. H. Scheike. Comment: Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Serie B*, 69:539–541, 2007.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18 (3):1269–1283, 1990.

S. Murphy and A. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95:449–485, 2000.

S. A. Murphy. Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22(2):712–731, 1994.

W. K. Newey. Semiparametric efficient bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics (The Harold Cramér Volume)*, pages 416–444. Almquist and Wiksells, Uppsala, Sweden, 1959.

S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787, 2018.

P. Riviere, C. Tokeshi, J. Hou, V. Nalawade, R. Sarkar, A. Paravati, M. Schiaffino, B. Rose, R. Xu, and J. Murphy. Claims-based approach to predict cause-specific survival in men with prostate cancer. *JCO Clinical Cancer Informatics*, 3:1–7, 2019.

J. Robins, S. D. Mark, and W. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

J. M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second Seattle Symposium in Biostatistics*, pages 189–326, New York, 2004. Springer.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

J. M. Robins and A. Rotnitzky. Comment on "inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

I. Sason. On refined versions of the Azuma-Hoeffding inequality with applications in information theory. *ArXiv e-prints:1704.07989*, 2013.

R. Satkunasivam, A. E. Kim, M. Desai, M. M. Nguyen, D. I. Quinn, L. Ballas, J. P. Lewinger, M. C. Stern, A. S. Hamilton, M. Aron, and I. S. Gill. Radical prostatectomy or external beam radiation therapy vs no local therapy for survival benefit in metastatic prostate cancer: A seer-medicare analysis. *The Journal of Urology*, 194:378–385, 2015.

T. A. Severini and W. H. Wong. Profile likelihood and conditionally parametric models. *Annals of Statistics*, 20:1768–1802, 1992.

S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

Z. Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv*, 2018.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

S. A. van de Geer. The deterministic Lasso. In *Joint Statistical Meeting proceedings*. American Statistical Association, 2007.

S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2): 614–645, 2008.

S. Vansteelandt and M. Joffe. Structural nested models and g-estimationa: The partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Y. Wang, M. Lee, P. Liu, L. Shi, Z. Yu, Y. A. Awad, A. Zanobetti, and J. D. Schwarts. Doubly robust additive hazards models to estimate effects of a continuous exposure on survival. *Epidemiology*, 28(6): 771–779, 2017.

D. Westreich and S. R. Cole. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677, 2010.

A. Ying, R. Xu, and J. Murphy. Two-stage residual inclusion for survival data and competing risks - an instrumental variable approach with application to SEER-Medicare linked data. *Statistics in Medicine*, 38(10):early view, 2019.

Y. Yu, J. Bradic, and R. J. Samworth. Confidence intervals for high-dimensional Cox models. *Statistica Sinica*, page to appear, 2019.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Serie B*, 76(1):217–242, 2014.

H. Zhang, L. Sun, Y. Zhou, and J. Huang. Oracle inequalities and selection consistency for weighted LASSO in high-dimensional additive hazards model. *Statistica Sinica*, 27:1903–1920, 2017.

M. Zhang and D. E. Schaubel. Contrasting treatment-specific survival using double-robust estimators. *Statistics in Medicine*, 31(30):4255–4268, 2012.

Q. Zhao. Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, 47(2): 965–993, 2019.

Y. Q. Zhao, D. Zeng, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 2015.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.