UNIVERSITY OF CALIFORNIA SAN DIEGO

Survival Analysis and Causal Inference: from Marginal Structural Cox to Additive Hazards Model and beyond

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Denise Rava

Committee in charge:

Professor Jelena Bradic, Chair
Professor Ronghui Xu, Co-Chair
Professor Loki Natarajan
Professor Dimitris Politis
Professor Armin Schwartzman

2021

The dissertation of Denise Rava is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents, my number one fans.

To Martina, Giorgio and Francesco,

my life's biggest blessing and my chosen family.

To Vasco, the love of my life.

EPIGRAPH

Quando realizzi un sogno, senti subito il bisogno di sognare ancora

—Vasco Rossi

Quando arrivano i conti sai, ognuno paga comunque i suoi, e stai tranquillo che io, i soldi ce li ho

—Vasco Rossi

Sai, essere libero costa soltanto qualche rimpianto

—Vasco Rossi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my advisors Professor Jelena Bradic and Professor Ronghui (Lily) Xu for their support. Professor Jelena Bradic first introduced me to the world of research and she taught me to be a curious, and tenacious researcher. Professor Ronghui (Lily) Xu taught me to be a careful, precise, and critical researcher. She is a role model and she has been an immense source of both academic and personal support.

I need to thank Professor Dimitris Politis, Professor Todd Kemp, Professor David Quarfoot and Professor Bruce Driver for teaching me, for working with me and for believing in me.

Thanks to all the friends I have shared my California life with, Holly, Giselle, Ayesha, Saribel, Ana, and Tere, for introducing me to their cultures, for the memories and for making this place a second home for me. Thanks to Pietro and Davide for struggling with me throughout this Ph.D. journey and for bringing a little bit of Italy over here. Holly and Ayesha truly deserve more than just a thank you. In the past year they have been there for me both in good and bad times, supporting me when I needed the most.

A very special thanks goes to West, for always showing me the light when all I could see was the dark.

Thanks to my mum and dad and my chosen family, Martina, Giorgio, and Francesco. My number one fans, they have been supporting me from Italy without, never once, making me feel too far away. They have always believed in me to the point that I have started to believe in myself. They have always made me think I have something to come home for. They are my biggest accomplishment, my roots, and my home and I could not live without their unconditional love.

Last but not least thanks to Vasco that has been bringing joy, motivation, and emotions into my life for years. His ’In bocca al lupo’ screamed at the end of ’Vita spericolata’ have carried me

all the way here and his ’favola di Cenerentola’ has made me believe I can also have my own.

Chapter 1, in full, is a reprint of the material as it appears in Statistics in Medicine. Rava, Denise; Xu, Ronghui. Explained variation under the additive hazards model, 40.1:85-100,2021. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it may appear in Lifetime Data Analysis, 2021, Rava, Denise; Bradic, Jelena. The dissertation author was the primary investigator and author of this paper.

Chapter 3, is currently being prepared for submission for publication of the material. Rava, Denise; Xu, Ronghui. The dissertation author was the primary investigator and author of this material.

Chapter 4, is currently being prepared for submission for publication of the material. Rava, Denise; Bradic, Jelena; Xu, Ronghui. The dissertation author was the primary investigator and author of this material.

VITA

2013	Laurea Triennale in Mathematics *cum laude*, Università degli Studi dell'Insubria, Como
2015	Laurea Magistrale in Mathematics *cum laude*, Università degli Studi dell'Insubria, Como
2016-2021	Graduate Teaching Assistant, University of California, San Diego
2021	Ph. D. in Mathematics with a Specialization in Statistics, University of California, San Diego

PUBLICATIONS

Rava Denise and Xu Ronghui, "Explained variation under the additive hazards model", *Statistics in Medicine*, 40.1: 85-100, 2021.

Rava Denise and Jelena Bradic, "DeepHazard: neural network for time-varying risks", *Lifetime Data Analysis*, Under Review, 2021

ABSTRACT OF THE DISSERTATION

Survival Analysis and Causal Inference: from Marginal Structural Cox to Additive Hazards Model and beyond

by

Denise Rava

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2021

Professor Jelena Bradic, Chair
Professor Ronghui Xu, Co-Chair

In Chapter 1 we study explained variation under the additive hazards regression model for right-censored data. We consider different approaches for developing such a measure, and focus on one that estimates the proportion of variation in the failure time explained by the covariates. We study the properties of the measure both analytically, and through extensive simulations. We apply the measure to a well-known survival dataset as well as the linked surveillance, epidemiology, and end results-Medicare database for prediction of mortality in early stage prostate cancer patients using high-dimensional claims codes.

In Chapter 2 we propose a new flexible method for survival prediction: DeepHazard, a neural network for time-varying risks. Prognostic models in survival analysis are aimed at understanding the relationship between patients covariates and the distribution of survival time. Traditionally, semiparametric models, such as the Cox model, have been assumed. These often rely on strong proportionality assumptions of the hazard that might be violated in practice. Moreover, they do not often include covariates' information updated over time. Our approach is tailored for a wide range of continuous hazards forms, with the only restriction of being additive in time. A flexible implementation, allowing different optimization methods, along with any norm penalty, is developed. Numerical examples illustrate that our approach outperforms existing state-of-the-art methodology in terms of predictive capability evaluated through the C-index metric. The same is revealed on the popular real datasets as METABRIC, GBSG, ACTG and PBC.

In Chapter 3 we consider the conditional treatment effect for competing risks data in observational studies. While it is described as a constant difference between the hazard functions given the covariates, we do not assume the additive hazards model in order to adjust for the covariates. We derive the efficient score for the treatment effect using modern semiparametric theory, as well as two doubly robust scores with respect to both the assumed propensity score for treatment and the censoring model, and the outcome models for the competing risks. We provide the asymptotic distributions of the estimators when the two sets of working models are both correct, or when only one of them is correct. We study the inference based on these estimators using simulations. The estimators are applied to the data from a cohort of Japanese men in Hawaii followed since 1960s in order to study the effect of mid-life drinking behavior on late life cognitive outcomes.

In Chapter 4 we consider doubly robust estimation of the causal hazard ratio in observational studies. The treatment effect of interest, described as the constant ratio between the hazard functions of the two potential outcomes, is parametrized by the Marginal Structural Cox Model. Under the assumption of no unmeasured confounders, causal methods, as Cox-IPW, have been developed for estimation of the treatment effect of interest. However no doubly robust methods have been proposed under the Marginal Structural Cox model. We develop an AIPW estimator for this popular model that is both model and rate-doubly robust with respect to the treatment assignment model and the conditional outcome model. The proposed estimator is applied to the data from a cohort of Japanese men in Hawaii followed since 1960s in order to study the effect of mid-life alcohol exposure on overall death.

Chapter 1

Explained Variation under the Additive Hazards Model

1.1 Introduction

The additive hazards model (Aalen, 1980, 1989) has received increasing attention lately for the analysis of censored survival data. It is not just an alternative to the more widely used Cox model when the proportional hazards assumption is violated; it has also been argued to be more suitable for causal inferences in estimating treatment effects because the Cox model is not collapsible (Martinussen and Vansteelandt, 2013). In contrast, the additive hazards model behaves mostly like a linear model including collapsibility, in the sense that one can integrate out an independent covariate from the model and still end up with an additive hazards model, with the same regression coefficients for all the other covariates. For this reason it has been used in the development of instrumental variable approaches for survival data including competing risks (Tchetgen Tchetgen et al., 2015; Li et al., 2015; Zheng et al., 2017; Brueckner et al., 2019; Ying et al., 2019). The collapsibility as well as other behaviors similar to a linear model, has also enabled the additive hazards model to be used in mediation analysis of survival data (Fosen et al., 2006; Martinussen, 2010; Martinussen et al., 2011; VanderWeele, 2013; Aalen et al., 2020). In addition, doubly robust methods have been developed for estimating treatment effects and applied in practice under the additive hazards model including for optimal treatment regimes (Wang et al., 2017; Kang et al., 2018; Blomberg et al., 2019), while the noncollapsibility of the Cox model presents an obstacle in the development of doubly robust method when confounders are present (Dukes et al., 2019a).

Estimation and inference procedures have been well developed and implemented under the additive hazards model (eg. R package 'timereg'), and diagnostic methods have also been proposed (Yuen and Burke, 1997; Kim and Lee, 1998; Scheike and Martinussen, 2006). However, another important aspect as the model becomes more widely used, is explained variation or measures of predictability, often referred to as R^2. O'Quigley and Xu (2012) provide detailed illustrations of how such measures are used to evaluate the clinical importance of prognostic factors. Müller et al. (2008) and Hielscher et al. (2010) explored the use of R^2 measures in genetic studies to quantify the impact of genetic variants or high dimensional gene expression on survival phenotypes, while Preseley et al. (2011) applied them to surrogate evaluation. Very recently applications of measure of dependence to ultrahigh dimensional variable screening were explored in Kong et al. (2019). In the context where the estimation of treatment effect is of primary concern, following the fit of the additive hazards models it is also natural to provide estimates of predicted survival given the covariates (Ying et al., 2019). However, measures of explained variation have not been examined under the additive hazards model to our best knowledge.

Explained variation has been well studied in the literature under the Cox regression model for right-censored data. Kent and O'Quigley (1988) first defined a measure of dependence for censored survival data, making use of the Kullback-Leibler information gain. It is based on the conditional distribution of the time to event random variable T given the covariates Z. A later work by Xu and O'Quigley (1999b) considered instead the conditional distribution of Z given T, using also the information gain. This latter measure can be readily extended to time-dependent covariates. A simple approximation to this second measure was described in O'Quigley et al. (2005), which can be easily computed using the partial likelihood ratio test statistic following the fit of the Cox model. Preseley et al. (2011) advocated for these information gain based measures.

Another approach to defining explained variation makes use of the residuals. This originated from the R^2 under the linear regression model, which can be written as one minus the ratio of the residual sum of squares over the total sum of squares. It is also well-known that these two sums of squares estimate the residual variance and the total variance, respectively. O'Quigley and Flandre (1994) proposed to use the Schoenfeld residuals under the Cox model, in a similar way to the R^2 under linear regression. It has been shown that when the Cox model appears to be a reasonably fit to the data, this measure and the one above based on information gain, tend to give comparable quantifications of explained variation (O'Quigley and Xu, 2012).

Other approaches have also been considered in the literature for right-censored data. Schemper and Kaider (1997) proposed to compute the correlation coefficients between the failure rankings and the covariates, using multiple imputation to handle the censored data. We note that inference under the Cox model is only based on the ranks of the failure times, hence nonparametric correlation coefficients like Kendall's tau or Spearman correlation might be considered. However, as it is known and we also elaborate below, inference under the additive hazards model is not rank based.

Finally and not restricted to the survival context, previous experiences in describing explained variation outside the classic linear model have also considered the direct decomposition of the total variance in the outcome, and quantifying the proportion that is explained by the covariates.

Depending on the model, this can sometimes be a straightforward approach, such as under the linear mixed effects model (Xu, 2003; Honerkamp-Smith and Xu, 2016), or under the accelerated failure time (AFT) models (Chan et al., 2018).

In this work we consider the semiparametric additive hazards model. We aim to quantify the explained variation under this model. It turns out that the last approach described above, i.e. the direct decomposition of the total variation into components of explained and unexplained (or residual) variation, is easily computable as well as interpretable under the additive hazards model. In the following we will first focus on its development, investigate its properties, and illustration how it might be used in practice to quantify the predictive power of a set of prognostic variables, and also for use in variable selection procedures. We will defer discussion to the end of the paper why some of the other approaches described above do not work under the additive hazards model.

The rest of the paper is organized as follows. After a review of the semiparametric additive hazards model and its inference in the next section, we describe explained variation and its estimator in section 1.3. In section 1.4, we study the properties of the measure, both the population and the sample-based versions. Section 1.5 further explores the behavior of the measures using simulation, under different censoring scenarios, different covariate distributions, different baseline hazard functions, and beyond. We apply the measure to real data sets in Section 1.6, and we conclude with discussion in the last section.

1.2 Semiparametric Additive Hazards Model

Let T be the failure time random variable of interest, Z be a vector of covariates, and C be the censoring time random variable. Let $X = \min(T, C)$ and $\delta = I(T \leq C)$ where $I(\cdot)$ is the indicator function. We observe a random sample (X_i, Z_i, δ_i), $i = 1, \ldots, n$. The semiparametric additive hazards model Lin and Ying (1994a) assumes that the conditional hazard function

$$\lambda(t|Z) = \lambda_0(t) + \beta^\top Z, \tag{1.1}$$

where $\lambda_0(t)$ is the baseline hazard and β is a vector of regression effects. We will also use the counting process notation: $N(t) = I\{X \leq t, \delta = 1\}$ and $Y(t) = I\{X \geq t\}$ are the counting process of events and the at-risk process, respectively.

Under model (1.1), an estimator for β was proposed by Lin and Ying (1994b):

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^\infty Y_i(t) \{Z_i - \bar{Z}(t)\}^{\otimes 2} dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^\infty \{Z_i - \bar{Z}(t)\} dN_i(t)\right], \tag{1.2}$$

where $\bar{Z}(t) = \sum_{i=1}^n Y_i(t) Z_i / \sum_{i=1}^n Y_i(t)$. We note that unlike under the Cox model, the above estimator of β is not rank based in that it depends on the values of X_i's beyond their ranks in the data set. It can be shown that, if $g(\cdot)$ is a strictly increasing function, then $g(T)$ in general no longer follows a semiparametric additive hazards model. In the special case where g is multiplication by a constant $c > 0$, then $\tilde{T} = cT$ still follows a semiparametric additive hazards model, but the regression coefficient is rescaled by c: $\tilde{\beta} = \beta / c$.

The cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is estimated by

$$\tilde{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n \left(dN_i(u) - Y_i(u)\hat{\beta}(u)^\top Z_i du\right)}{\sum_{j=1}^n Y_j(u)}. \tag{1.3}$$

In the following we write out the integral in (1.3), which is not a step function. Denote the K ordered distinct observed failure times $t_1 < \ldots < t_K$. We have for $k = 1, \ldots, K$:

$$\tilde{\Lambda}_0(t_k) = \sum_{l=1}^{k} \frac{\delta_l d_l}{r_l} - \sum_{l=1}^{k} \hat{\beta}^\top \bar{Z}(t_l)(t_l - t_{l-1}), \tag{1.4}$$

where d_l and r_l are the number of events and number at risk at time t_l, respectively. In addition, for any $t_k \leq t < t_{k+1}$,

$$\tilde{\Lambda}_0(t) = \sum_{l=1}^{k} \frac{\delta_l d_l}{r_l} - \sum_{l=1}^{k} \hat{\beta}^\top \bar{Z}(t_l)(t_l - t_{l-1}) - \hat{\beta}^\top \bar{Z}(t_{k+1})(t - t_k). \tag{1.5}$$

The resulting estimated survival function $\tilde{S}(t|z) = \exp(-\tilde{\Lambda}_0(t) - \hat{\beta}\top zt)$ is not guaranteed to be non-increasing; therefore we make use of the following adjusted version Lin and Ying (1994a): $\hat{S}(t|z) = \min_{s \leq t}\left\{\tilde{S}(s|z)\right\}$. The adjusted version $\hat{S}$ is asymptotically equivalent to $\tilde{S}$, and the process $\sqrt{n}(\hat{S}(\cdot|z) - S(\cdot|z))$ converges wealy to a zero-mean Gaussian process Lin and Ying (1994a). We note that taking minimum over $s \leq t$ leads to no closed-form expression and the quantity needs to be computed numerically. However, it is imperative that we work with a proper distribution or equivalently, survival, function, in order to estimate the moments below.

1.3 Explained Variation

The explained variation, as described in the survival context by O'Quigley and Xu (2012), can be defined as

$$\Omega^2 = 1 - \frac{E\{\mathrm{Var}(T \mid Z)\}}{\mathrm{Var}(T)} = \frac{\mathrm{Var}\{E(T \mid Z)\}}{\mathrm{Var}(T)}. \tag{1.6}$$

This is consistent with the regression setting of model (1.1) for the conditional distribution of T given Z, as the proportion of variation of T explained by Z out of the total variation of T. As pointed out in O'Quigley (2008) page 33, by virtue of the Chebyshev-Bienayme inequality, the variance can be seen as a measure of predictability, and therefore the explained variation may also have an interpretation as predictability.

In practice for survival studies, there is often a finite upper bound of time τ due to administrative censoring, so that all the observable data are conditional upon $T < \tau$. We then define

$$\Omega_\tau^2 = 1 - \frac{E\{\mathrm{Var}(T \mid Z, T < \tau)\}}{\mathrm{Var}(T \mid T < \tau)} = \frac{\mathrm{Var}\{E(T \mid Z, T < \tau)\}}{\mathrm{Var}(T \mid T < \tau)}. \tag{1.7}$$

Obviously when there is no censoring, $\Omega^2 = \Omega_\infty^2$; and in the following for uniformity of notation, we allow $\tau \leq \infty$.

We can estimate directly the quantities in (1.7) under model (1.1). To estimate $E\{\mathrm{Var}(T \mid Z, T < \tau)\}$ or $\mathrm{Var}\{E(T \mid Z, T < \tau)\}$, we first integrate with respect to an estimated distribution of T given Z and $T < \tau$:

$$\hat{S}(t \mid Z, T < \tau) = \frac{\hat{S}(t \mid Z) - \hat{S}(t_K \mid Z)}{1 - \hat{S}(t_K \mid Z)} I\{t \leq t_K\} \tag{1.8}$$

We then integrate with respect to $\mathbb{P}_n$, the empirical distribution of Z. Denote the resulting estimates $E_n\left\{\widehat{\mathrm{Var}}(T \mid Z, T < \tau)\right\}$ and $\mathrm{Var}_n\left\{\hat{E}(T \mid Z, T < \tau)\right\}$, respectively. For example,

$$E_n\left\{\widehat{\mathrm{Var}}(T \mid Z, T < \tau)\right\} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{E}\left(T^2 \mid Z_i, T < \tau\right) - \left\{\hat{E}\left(T \mid Z_i, T < \tau\right)\right\}^2\right], \tag{1.9}$$

where the expressions for the quantities in the right-hand side above are given later in the section.

To estimate $\mathrm{Var}(T \mid T \leq \tau)$, we can use

$$\widehat{\mathrm{Var}}(T \mid T < \tau) = \hat{E}\left(T^2 \mid T < \tau\right) - \left\{\hat{E}\left(T \mid T < \tau\right)\right\}^2. \tag{1.10}$$

In order to estimate the marginal survival function, we may use the nonparametric Kaplan-Meier

(KM) estimator. Alternatively, we may use:

$$\hat{S}(t \mid T < \tau) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}(t \mid Z_i, T < \tau). \tag{1.11}$$

It can be shown that, if (1.11) is used in estimating the expectations in (1.10), then we have the following decomposition:

$$\widehat{\mathrm{Var}}(T \mid T < \tau) = E_n\left\{\widehat{\mathrm{Var}}\,(T \mid Z, T < \tau)\right\} + \mathrm{Var}_n\left\{\hat{E}(T \mid Z, T < \tau)\right\}. \tag{1.12}$$

Combining all of the above, we obtain R^2_τ as a consistent estimator of Ω^2_τ under model (1.1):

$$R^2_\tau = 1 - \frac{E_n\left\{\widehat{\mathrm{Var}}\,(T \mid Z, T < \tau)\right\}}{\widehat{\mathrm{Var}}\,(T \mid T < \tau)} = \frac{\mathrm{Var}_n\left\{\hat{E}\,(T \mid Z, T < \tau)\right\}}{\widehat{\mathrm{Var}}\,(T \mid T < \tau)}. \tag{1.13}$$

We also denote $R^2 = R^2_\infty$ when $\tau = \infty$.

Finally, to compute the quantities in (1.13), we have:

$$\begin{aligned} \hat{E}\,(T \mid z, T < \tau) &= \int_0^\tau \hat{S}\,(t \mid z, T < \tau)\,dt \\ &= \frac{1}{1 - \hat{S}\,(t_K \mid z)} \int_0^{t_K} \hat{S}\,(t \mid z)\,dt - \frac{1}{1 - \hat{S}\,(t_K \mid z)} \hat{S}\,(t_K \mid z)\,t_K, \end{aligned} \tag{1.14}$$

and

$$\begin{aligned} \hat{E}\left(T^2 \mid z, T < \tau\right) &= 2\int_0^\tau t \cdot \hat{S}(t \mid z, T < \tau)dt \\ &= \frac{2}{1 - \hat{S}\,(t_K \mid z)} \int_0^{t_K} t\hat{S}\,(t \mid z)\,dt - \frac{1}{1 - \hat{S}\,(t_K \mid z)} \hat{S}\,(t_K \mid z)\,t_K^2. \end{aligned} \tag{1.15}$$

Since there is no closed-form expression for $\hat{S}(t \mid Z)$, the integrals in the above are computed using the trapezoidal rule. We partition the interval $[0,\tau]$ first using $t_1,\ldots,t_K$; additional points are added to create a grid no wider than 0.01 between two adjacent points. We then use an iterative halving process, i.e. adding the midpoints between any two adjacent points to the grid, until the change in the resulting R^2_τ is less than 0.01 in absolute value.

The quantities in $\widehat{\text{Var}}(T \mid T < \tau)$ can be computed in a similar fashion using (1.11).

1.4 Properties of Ω^2 and R^2

The desirable properties of a measure of explained variation are best understood under a linear regression model, including: 1) it lies between zero and one; 2) it takes the value zero when there is no regression effect; 3) it increases with the strength of the regression effect; 4) it tends to one as the regression effect tends to infinity; 5) it is invariant under certain transformations of the dependent and independent variables, depending on the model. For the last property, the transformation is linear under the linear regression model, and is rank-preserving for the failure time under the semiparametric Cox regression model O'Quigley and Xu (2012).

In the following we investigate if the above properties hold for the measures defined in the last section.

- The facts that $0 \le \Omega^2_\tau \le 1$ and $0 \le R^2_\tau \le 1$ follow immediately from their definitions (1.7) and (1.13), assuming that the latter is estimated using (1.11).

- When $\beta = 0$, $\Omega^2_\tau = 0$ because independence between T and Z implies that $\text{Var}\{E(T \mid T < \tau, Z)\} = \text{Var}\{E(T \mid T < \tau)\}$. Also $R^2_\tau = 0$ if it happens that the estimated coefficient $\hat{\beta} = 0$. Otherwise, the sample based measure $R^2_\tau > 0$, but is expected to be small since it is a consistent estimate of $\Omega^2_\tau = 0$.

- It is analytically difficulty to prove that Ω^2_τ increases with $|\beta|$ in general. However, for simpler settings such as a binary Z and $\tau = \infty$, we can prove it analytically and this is given in the Appendix. For more general settings, we illustrate this via simulation.

- It has been known that the quantity Ω^2 defined in (1.6) can be bounded strictly less than one O'Quigley and Xu (2012). For a binary Z, if we assume that $T \mid Z = 0$ has finite second moment, then we can show by the dominated convergence theorem that:

$$\lim_{\beta\to\infty} \Omega^2_\infty = 1 - \frac{\frac{1}{2}\left[2\int_0^\infty t\exp\{-\Lambda_0(t)\}dt - [\int_0^\infty \exp\{-\Lambda_0(t)\}dt]^2\right]}{\int_0^\infty t\exp\{-\Lambda_0(t)\}dt - \frac{1}{4}[\int_0^\infty \exp\{-\Lambda_0(t)\}dt]^2}. \tag{1.16}$$

  For example, when $\lambda_0(t) = 1$, $\lim_{\beta\to\infty}\Omega^2_\infty = 0.333$; and this is the exponential case discussed in O'Quigley and Xu (2012). When $\lambda_0(t) = t$, $\lim_{\beta\to\infty}\Omega^2_\infty = 0.647$; and when $\lambda_0(t) = 1/(2\sqrt{t})$, $\lim_{\beta\to\infty}\Omega^2_\infty = 0.091$. Similar calculation can be done for covariates with continuous distribution:

$$\lim_{\beta\to\infty} \Omega^2_\infty = 1 - \lim_{\beta\to\infty} \frac{\int_{\mathcal{Z}}\left[\int_0^\infty 2te^{-\Lambda_0(t)-\beta^\top Zt}dt - \left[\int_0^\infty e^{-\Lambda_0(t)-\beta^\top Zt}dt\right]^2\right]g(z)dz}{\int_{\mathcal{Z}}\int_0^\infty 2te^{-\Lambda_0(t)-\beta^\top Zt}dtg(z)dz - \left[\int_{\mathcal{Z}}\int_0^\infty e^{-\Lambda_0(t)-\beta^\top Zt}dtg(z)dz\right]^2}, \tag{1.17}$$

  where $g(Z)$ is the density of the covariates and $\mathcal{Z}$ is their sample space. This limit may not be equal to one and it depends on the form of $\lambda_0(t)$ and the distribution of Z; for example, when $Z \sim U\left[0,\sqrt{3}\right]$ and $\lambda_0(t) = 1$, $\lim_{\beta\to\infty}\Omega^2_\infty = 0.500$.

- By their definitions and simple algebra, it can be shown that Ω^2_τ and R^2_τ are invariant under linear transformations of Z and when T is rescaled by a positive constant.

In summary, we have the following properties:

1) $0 \le \Omega^2_\tau \le 1$, and $0 \le R^2_\tau \le 1$;

2) $\Omega^2_\tau = 0$ when $\beta = 0$, and $R^2_\tau = 0$ if $\hat{\beta} = 0$;

3) Ω^2_τ increases with $|\beta|$;

4) Ω^2_τ and R^2_τ are invariant under any linear transformation of Z and rescaling of T.

1.5 Simulations

In the following we further study the properties of the measures through simulations. In addition to the properties mentioned above, we also investigate: 1) the effect of baseline hazard on explained variation; 2) explained variation under nested models. As we have more experience with explained variation under the Cox proportional hazards regression model, we also investigate 3) how the measure compares with a similar one under the Cox model, when both models are valid; and 4) explained variation of Z give T, which has been advocated for use under the Cox model.

All simulations below were carried out with sample size 1000, and 100 simulation runs each. All the results are reported as mean with standard deviation (SD) over the simulation runs in $(\cdot)$. As the simulation has been extensive, we have chosen to display the representative scenarios that carry meaningful messages, as opposed to every combination of all possible parameters and settings.

1.5.1 Basic properties

As $|\beta|$ increases

We first simulated with $\lambda_0(t) = 1$ and different values β =1, 3, 15 and 50, Z from Uniform $[0, \sqrt{3}]$ as well as binary 0,1 with equal probabilities. Note that these two covariate distributions have the same variance 0.25, rendering the measures comparable for a given β value. The censoring distribution was uniform $[0,\tau]$. We computed the Ω^2_τ values as follows. When there was no censoring we computed it analytically by definition using the fact that $T \sim$ Exponential $(1+\beta Z)$. When there was censoring, we took a single large sample size of 100,000, and used the R^2_τ value computed with the true β and the true λ_0 to approximate Ω^2_τ.

From Figure 1.1 and Table 1.1 we see that R^2_τ and Ω^2_τ values are close in all cases, both increasing with $|\beta|$ as expected. The effect of τ reflects different follow-up periods, which also leads to different amounts of censoring. It is seen that the patterns of change with τ is different depending on the distribution of Z. It is more pronounced with binary Z especially for that larger β values, likely because the censor percentages are much higher in that case.

Effect of $\lambda_0(\cdot)$

We consider here a binary Z taking values 0,1 with equal probabilities. We consider $\lambda_0(t) = 1, t$ and $1/(2\sqrt{t})$. In Figure 1.2 we plot the density of T for each group, to show how the two groups differ in each scenario. The mean of R^2_∞ over the 100 simulations are printed on each configuration. From Figure 1.2 we see that the R^2_∞ values tend to be larger when the two groups indexed by $Z = 0$, 1 have different concentrations of failure times, i.e. different shapes of the density functions, such as in the case of $\lambda_0(t) = t$. On the contrary, with $\lambda_0(t) = 1/(2\sqrt{t})$ the two density functions have very similar shapes, resulting much smaller R^2_∞ values. As noted earlier, the upper bound of Ω^2 for the three cases are 0.091, 0.333 and 0.647, respectively.

Nested models

Next we consider a limited set of simulations with data generated under $\lambda(t|Z) = \lambda_0(t) + Z_1 + 3Z_2 + Z_3$, where the covariates Z_1, Z_2 and Z_3 were independently drawn from Uniform $[0,\sqrt{3}]$ and the baseline hazard was in turn equal to $1, t$ and $1/(2\sqrt{t})$. We also consider an additional pure noise covariate $Z_4 \sim$ Uniform $[0,\sqrt{3}]$, not used in the data generating mechanism. We consider the following models listed in Table 1.2: three univariate models with each of Z_1, Z_2 and Z_3, respectively; a model with only Z_1 and Z_3; a model with all the three Z_1, Z_2, Z_3; and a model with the three covariates plus the pure noise Z_4. We see from Table 1.2 that R^2_∞ increases with the complexity of the models: R^2_∞ with both Z_1 and Z_3 is larger than with Z_1 or Z_3 alone; meanwhile, since Z_2 has a strong effect as reflected in its regression coefficient, R^2_∞ with Z_2 alone is larger than with both Z_1 and Z_3. The measure is substantially larger with all three covariates Z_1, Z_2 and Z_3 than under any of the previous models. With the noise variable Z_4 added to the model, R^2_∞ increases very slightly from 0.122 to 0.124, for example. This also informs us how to use the R^2 type measures for model selection: if the addition of a variable only increases the R^2 very slightly, it is perhaps not worth the cost of an extra degree of freedom. This is consistent with the concept of adjusted R^2, which explicit adjusts for the number of degrees of freedom. We further discuss this in the applications later.

1.5.2 Comparison with the measure under the Cox Model

As discussed earlier the semiparametric additive hazards model behaves somewhat differently from the semiparametric Cox model. Here we compare R^2_τ as defined in (1.13) under the two models when both models are valid. We consider a binary Z and constant baseline hazard; this is a case where both the semiparametric additive hazards model (1.1) and the classic Cox model hold.

Under the Cox model $S(t \mid Z) = \{S_0(t)\}^{\exp(\beta Z)}$, where the regression parameter is typically estimated using the partial likelihood, and the baseline survival function via the Breslow's estimate of the cumulative baseline hazard. We can then similarly estimate the explained variation as defined in (1.6) or (1.7), using a similar approach as described in Section 3. We denote this as R^2_{cox}. Both R^2_{cox} and R^2_τ thus defined should be consistent for the same Ω^2_τ. In Table 1.3 we again simulated with $\lambda(t|Z) = 1 + \beta Z$ for a binary Z, $\beta = 1, 3, 15$ and 50, with no censoring or 30% censoring . As expected, the values of R^2_{cox} and R^2_τ are indeed very close to each other.

1.5.3 Explained variation of Z given T

O'Quigley and Xu (2012) advocated for considering the explained variation of Z given T under the Cox regression model. One main advantage of this approach is that the resulting measure tend not to be bounded strictly less than one. In addition, considering Z given T is also consistent with the sequential conditioning and counting process notation often used in survival analysis. Following O'Quigley and Flandre (1994) and O'Quigley and Xu (2012), we consider in particular the covariate residual (also called Schoenfeld residual under the Cox model) based approach.

In order to obtain the residuals of Z, we need to estimate the conditional distribution of Z given T. A theorem from Xu and O'Quigley (1999b,a) can be readily adapted to provide a consistent estimate of this conditional distribution under model (1.1):

Theorem 1. *Under model* (1.1) *and independent censoring, assuming that* $\lambda_0(t)$ *is known (or otherwise consistently estimated), the conditional distribution of Z given T is consistently estimated by*

$$\hat{P}(Z \le z \mid T = t) = \frac{\sum_{Z_j \le z} Y_j(t)\left(\lambda_0(t) + \hat{\beta}^T Z_j\right)}{\sum_{l=1}^{n} Y_l(t)\left(\lambda_0(t) + \hat{\beta}^T Z_l\right)}. \tag{1.18}$$

The proof of the above theorem is similar to that of Theorem 1 in Xu and O'Quigley (1999b,a) but applied to model (1.1).

In practice $\lambda_0(t)$ is unknown, and also not readily estimated by the typical software that fit the additive hazards model. Our investigation here is of exploratory nature, aimed to understand the behaviors of the explained variation of T give Z versus Z given T. In simulations below we use the true $\lambda_0(t)$. Denote

$$\hat{E}_\beta\left(Z \mid t\right) = \frac{\sum_{j=1}^{n} Z_j Y_j(t)\left(\lambda_0(t) + \beta Z_j\right)}{\sum_{l=1}^{n} Y_l(t)\left(\lambda_0(t) + \beta Z_l\right)}. \tag{1.19}$$

The residuals under the fitted model and under the 'null' model where $\beta = 0$ are, respectively:

$$r_i(\hat{\beta}) = Z_i - \hat{E}_{\hat{\beta}}\left(Z \mid X_i\right), \; r_i(0) = Z_i - \hat{E}_0\left(Z \mid X_i\right), \tag{1.20}$$

where $E_0\left(Z \mid X_i\right)$ is simply the empirical average of Z in the risk set at time X_i. Therefore for a scalar Z we may define

$$R^2_{Z|T} = 1 - \frac{\sum_{i=1}^{n} r_i^2(\hat{\beta})}{\sum_{i=1}^{n} r_i^2(0)}.$$

The extension to multivariate Z was described in O'Quigley and Xu (2012) and can be easily adopted here.

We simulated under $\lambda(t) = 1 + \beta Z$, with a binary Z and equal probabilities of 0, 1. In Table 1.4 we see that unlike R^2_τ, the values of $R^2_{Z|T}$ approach one with increasing $|\beta|$. We further discuss the unknown $\lambda_0(t)$ in the last section.

1.6 Applications

1.6.1 Leukimia: FREIREICH DATA

We first apply the measure of explained variation to the Freireich et al. (1963) data, which consist of the remission times of 42 Leukimia patients in a randomized clinical trial treated with the drug 6-mercaptopurine (6-MP) versus placebo. The data set has been well-known in the survival analysis literature, and was in the first table of Cox and Oakes (1984). As a diagnostic plot in Figure

1.3 we show the difference of the cumulative hazard functions between the two treatment groups; under the semiparametric additive hazards model (1.1) this difference should be linear in time. From the figure we see that except for random noise due to limited sample size the difference shows a very nice linear trend, indicating that the semiparametric model (1.1) fits the data reasonably well. We note that in the R package ‘timereg’ that we used to fit the semiparametric additive hazards model, no diagnostic tools appear to be provided for checking this model.

We calculated $R^2 = 0.201$, indicating, as is known, good separation between the two groups' survival times. Typically if a single predictor, in particular a binary one, turns out to have an R^2 of around 20% say, it is considered to be a strong predictor. Previously the explained variation of Z given T under the Cox regression model had been calculated to be around 0.40 (ranging from 0.38 to 0.42 depending on the measure used) O'Quigley and Xu (2012). The Freireich data appears to be a data set that fits both the proportional hazards model and the additive hazards model reasonably well. Based on the simulation results, when the data fits both models, the explained variation of T given Z would be very close under the two models. The discrepancy between the R^2 values seen above are most likely attributable to the difference between the explained variation of Z given T and that of T given Z, as also illustrated in the simulations. In this case they otherwise reflect somewhat comparable strengths of association in our opinion.

1.6.2 Prostate cancer: SEER-MEDICARE DATA

We study the time to death of 29,657 prostate cancer patients with localized non-metastatic disease identified from the linked Surveillance, Epidemiology, and End Results (SEER) - Medicare database, diagnosed between 2004 and 2009. Following Hou et al. (2018) we consider the clinical and the demographical variables, plus the binary insurance claims codes from Medicare. The latter captures medical diagnoses and procedures through Healthcare Common Procedure Coding System

(HCPCS) codes, international classification of diseases (ICD)-9 diagnosis and procedure codes, etc. Each insurance claims code variable takes value one if that claim appeared within one year before diagnosis, and zero otherwise. Out of the 29,657 patients 3,543 died by the end of the follow-up which was December 2013 when the data were exported from the linked database.

The high dimensional data analysis of Hou et al. (2018) selected 143 variables to predict non-cancer mortality, and 9 variables to predict cancer mortality, in the context of these two competing risks. The same sets of variables were used in Riviere et al. (2019) and a complete list can be found in Table 1 and 2 of their supplemental material. For our analysis of explained variation, we combined these two sets of predictor for overall survival, which resulted in 146 variables: PSA, Gleason Score, age, race (black versus other), marital status (married versus other) and registry (California versus other), plus the claims codes. A table with the distributions of these variables can be found in the Supplemental Materials.

In Figure 1.5 we plot the difference of the cumulative hazard functions between groups as we did for the Freireich data above, to check the additive hazards model assumption. These are illustrated for six binary variables, the three demographical variables plus three claims codes that are not too sparse to plot. The plots indicate that the model seems to fit the data reasonably well.

We consider three models here. We first fit the data to the semiparametric additive hazards model with only the cancer-related clinical variables PSA and Gleason Score. We then add the four demographical variables. Finally we added the set of claim codes. The model fits are provided in the tables of the Supplement Materials. Table 1.5 summarizes the R^2 values obtained under these three models. In the first column of the table we see that the cancer-related clinical variables alone do not explain much (under 1%) variation in overall survival. This can at least be partially understood since only 734 out of the 3,543 total deaths in this data set were due to cancer. Demographical variables, on the other hand, do explain a substantial amount of variation in overall survival. This amount of explained variation is further increased, by a non-trivial amount, after adding in the claims codes previously identified from the high-dimensional SEER-Medicare database.

When high dimensional claims codes are used in the data analysis, there is often the concern of model over-fitting. In our case, with 3,543 death events and 146 total regressors, this may not be an issue. Nonetheless, we proceed to divide the data set randomly into two parts, a training set with 14,828 observations containing 1,803 deaths, and a test set with 14,829 observations containing 1740 deaths. We fit the additive hazards model to the training data set and obtain the estimates $\hat{\beta}$ and $\hat{\Lambda}_0(t)$. We use them to compute $\hat{S}(t|Z)$ on the test data set, and obtain an out-of-sample R^2_{out}. Such out-of-sample R^2 measures are often used in machine learning applications (eg. deep learning) in order to reduce the risk of overfitting. We report the R^2_{out} in Table 1.5. It is seen that, for this data, the R^2_{out} values are in fact slightly higher than the R^2 computed on the full data set, or the R^2_{train} computed on the training data set. Were there over-fitting, the R^2_{out} values would have been substantially lower. The discrepancy among the three quantities currently seen is mostly due to variability in the estimation of the conditional survival function and consequently of the total and explained variances. For comparison purposes, we also provide in the Supplemental Materials the three model fits to the training data set. We can compare the estimated coefficients with those using the full data set, and observe that the estimates for the statistically signficant ones are stable across the training versus the full data set.

At the suggestion of a reviewer, we compute the adjusted R^2, $R^2_{adj} = 1 - (1 - R^2)(n - 1)/(n - p - 1)$, for the three models. Here n is the sample size, and p is the number of the covariates included in the model. The R^2_{adj} is computed on the full data set. By definition $R^2_{adj} < R^2$, although no difference can be seen at three digits after the decimal point between the two measures for the first two models since p is so small compared to n. For the third model that includes 146 variables, the difference of 0.3% between the two does not appear to signify any over-fitting.

Finally we note that the explained variation of Z given T under the Cox model, denoted ρ^2, was calculated in Riviere et al. (2019) for this data. They computed $\rho^2 = 0.71$ for cancer mortality and $\rho^2 = 0.60$ for non-cancer mortality under competing risks setting. As discussed before, the numerical values of explained variation of T given Z are not directly comparable to those of Z given T. Considering that the former has an upper bound less than one, it is perhaps also within reasons to conclude that our analysis under the additive hazards model agrees with that of Riviere et al. (2019) about the contribution of the claims codes in explaining overall mortality for this prostate cancer patient population. This conclusion echoes the initial goal of the funded project that lead to the previous publications Hou et al. (2018); Riviere et al. (2019) to demonstrate that the high-dimensional insurance claims codes contain useful information about mortality in this patient population.

1.7 Discussion

In this paper we have studied explained variation under the semiparametric additive hazards model for right-censored survival data. The explained variation is shown to lie between zero and one, and to increase with the magnitude of the regression effect. It has been known, and is shown again here, that the explained variation of survival time given covariates can have an upper bound strictly less than one. Nonetheless, Ash and Shwartz (1999) argues convincingly that low R^2 values can be useful as a measure of model performance and prediction, and we have illustrated the same in our data analyses. Indeed in many of today's genome-wide association studies, polygenic risks scores are commonly assessed using R^2 measures, even though their values are typically very low (single digit of percentage points) for most diseases studied.

The semiparametric additive hazards model is different in several aspects from the historically more widely used semiparametric proportional hazards model. The model and hence its inference is not rank invariant, which makes it less familiar to most users in the seimparametric survival analysis field. This phenomenon also carries over to the explained variation under the model, leading to its dependence on the baseline hazard function. Of course, the choice of a model should depend on how close it is to the true data generating mechanism. On the other hand, as mentioned earlier the semiparametric additive hazards model is known to be collapsible, and this makes it more sensible to compare nested models which, as we have illustrated, is a common usage of R^2 type measures.

As reviewed in the Introduction, other approaches exist in the literature in order to develop R^2 type measures. In the Simulation section, we have considered a residual based approach, that relates to the explained variation of the covariates given the survival time. This was an approach advocated under the Cox proportional hazards model O'Quigley and Xu (2012), as it does not encounter the problem of being bounded strictly less than one. Unfortunately, for the additive hazards model, it requires the knowledge or consistent estimation of the baseline hazard function $\lambda_0(t)$, which is not provided in the commonly used software such as the R package 'timereg'. Smoothing methods such as kernels may be applied to $\widehat{\Lambda}_0(t)$, and can be potentially used here, but this is beyond the scope of this work. A third approach is based on information gain, but as it turns out, it also requires an estimate of $\lambda_0(t)$ under the additive hazards model.

The R package 'timereg' also allows β to vary with time, i.e. $\beta(t)$ in place of β in model (1.1). It estimates the cumulative $B(t) = \int_0^t \beta(u)du$, together with $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. It is possible to define an R^2 measure similar to what we have done in this paper; the computation is in fact simpler because the estimated conditional survival function $S(t|z)$ is a step function. To our best knowledge little experience exists in the literature to inform us when to use this more general nonparametric model versus the semiparametric model we have considered here. We have noticed that the nonparametric model does not appear suitable for the two data sets in this paper. The Freireich data set appears to have too small a sample size to the fit the nonparametric model, in that the resulting estimates are extremely bumpy and have large variation. The SEER-Medicare data set, on the other hand, is so sparse in the design matrix (i.e. many zero values for the claims codes), together with high percentage of censoring, that the resulting estimated $B(t)$ is practically constant zero. This is not difficult to see from the formula $\hat{B}(t) = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \int_0^t d\mathbf{N}(u)$, where $\mathbf{Z} = [Z_1, \ldots, Z_n]^\top$ and $\mathbf{N}(u) = [N_1(u), \ldots, N_n(u)]^\top$.

The R^2 measure of explained variation should not be confused with goodness-of-fit measures, although there are connections between these two concepts. Chauvel and O'Quigley (2017) show that the population version of the explained variation under the proportional hazards model will increase with improvements of fit, and that the best model from a large class of models maximizes the explained variation. They consider this in a similar setting as $\beta(t)$ in the above; see also Flander and O'Quigley (2019). However, due to issues in fitting $\beta(t)$ under the additive hazards model, we have not been able to observe a similar phenomenon. This would be worth future investigation once we are able to have a good estimate of $\beta(t)$, perhaps with smoothing techniques.

The R^2 measure developed in this work has been implemented in the R package 'R2Addhaz' and is publicly available on CRAN.

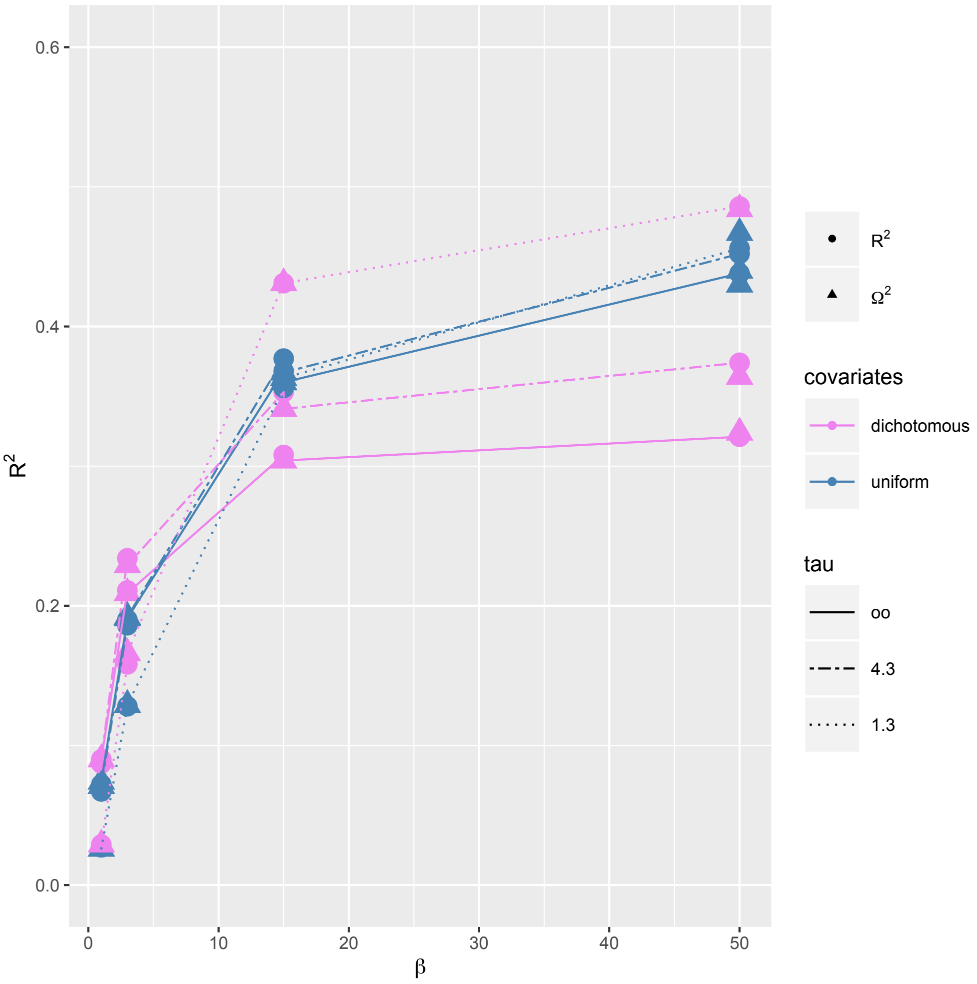


Figure 1.1: Ω^2_τ and R^2_τ values for different β and τ under the model $\lambda(t) = 1 + \beta Z$.

Table 1.1: Simulation results for different values of β and τ under the model $\lambda(t) = 1 + \beta Z$; in () are standard errors from simulation runs.

β	Z	τ	Censor	$\widehat{\beta}$	R^2_τ	Ω^2_τ
1	$U\left(0, \sqrt{3}\right)$	∞	0%	1.000 (0.117)	0.072 (0.017)	0.074
		4.3	14%	0.992 (0.128)	0.067 (0.017)	0.071
		1.3	39%	0.994 (0.151)	0.027 (0.013)	0.026
	Binary	∞	0%	1.000 (0.103)	0.090 (0.014)	0.090
		4.3	17%	1.001 (0.115)	0.087 (0.017)	0.090
		1.3	45%	1.006 (0.140)	0.029 (0.014)	0.029
3	$U\left(0, \sqrt{3}\right)$	∞	0%	2.996 (0.227)	0.190 (0.027)	0.191
		4.3	8%	2.984 (0.238)	0.186 (0.029)	0.192
		1.3	25%	2.997 (0.259)	0.128 (0.027)	0.129
	Binary	∞	0%	3.020 (0.184)	0.211 (0.018)	0.209
		4.3	14%	3.037 (0.231)	0.234 (0.020)	0.229
		1.3	38%	2.962 (0.210)	0.158 (0.033)	0.166
15	$U\left(0, \sqrt{3}\right)$	∞	0%	15.077 (0.765)	0.368 (0.044)	0.360
		4.3	3%	15.175 (0.916)	0.377 (0.047)	0.367
		1.3	10%	15.041 (0.925)	0.356 (0.050)	0.363
	Binary	∞	0%	15.053 (0.765)	0.308 (0.020)	0.304
		4.3	12%	14.943 (0.669)	0.353 (0.021)	0.341
		1.3	30%	15.083 (0.781)	0.431 (0.024)	0.431
50	$U\left(0, \sqrt{3}\right)$	∞	0%	49.438 (3.272)	0.438 (0.070)	0.430
		4.3	1%	49.878 (2.446)	0.452 (0.069)	0.440
		1.3	4%	49.741 (2.695)	0.456 (0.069)	0.467
	Binary	∞	0%	50.3373 (2.465)	0.321 (0.020)	0.324
		4.3	12%	49.761 (2.577)	0.374 (0.018)	0.364
		1.3	29%	50.056 (2.479)	0.486 (0.022)	0.484

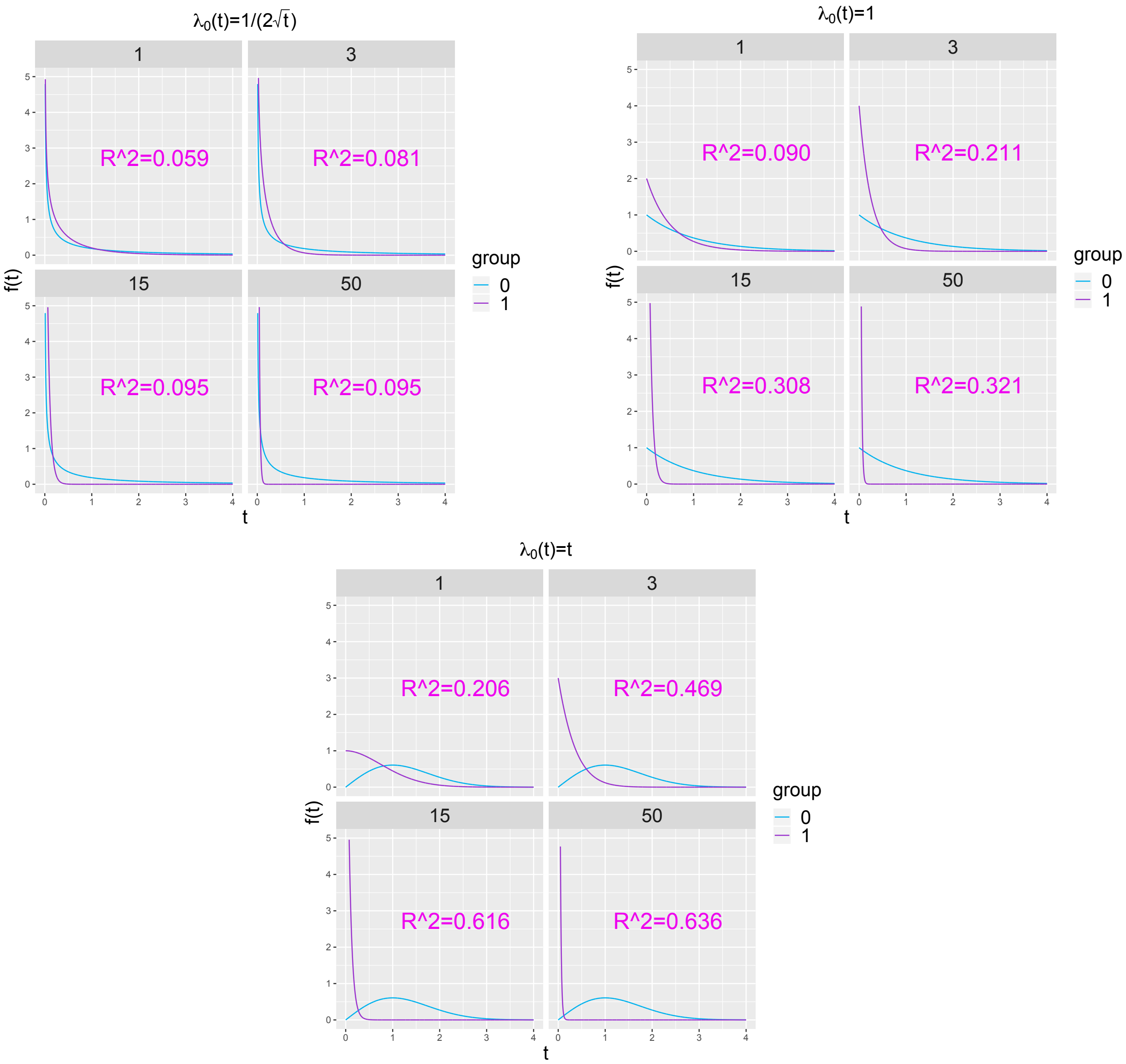


Figure 1.2: Density of T for each of $Z = 0$, 1 groups, superimposed with the average R^2 values over simulations for each configuration.

Table 1.2: R^2 values for nested models; in () are standard errors from simulation runs.

Model	$\lambda_0(t) = 1$	$\lambda_0(t) = t$	$\lambda_0(t) = 1/(2\sqrt{t})$
Z_1	0.012 (0.005)	0.016 (0.008)	0.008 (0.005)
Z_2	0.084 (0.019)	0.122 (0.020)	0.060 (0.015)
Z_3	0.013 (0.007)	0.015 (0.006)	0.008 (0.005)
Z_4	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)
$Z_1 + Z_3$	0.025 (0.010)	0.031 (0.011)	0.017 (0.010)
$Z_1 + Z_2 + Z_3$	0.122 (0.028)	0.174 (0.031)	0.087 (0.023)
$Z_1 + Z_2 + Z_3 + Z_4$	0.124 (0.029)	0.176 (0.031)	0.089 (0.024)

Table 1.3: Comparison of explained variation under the semiparametric additive hazards model and the semiparametric Cox model, when both models are correct; in () are standard errors from simulation runs.

β	Censor	R^2_τ	R^2_{cox}
1	0%	0.094 (0.015)	0.094 (0.015)
	30%	0.063 (0.018)	0.063 (0.016)
3	0%	0.208 (0.015)	0.208 (0.015)
	30%	0.207 (0.027)	0.208 (0.026)
15	0%	0.306 (0.022)	0.306 (0.022)
	30%	0.433 (0.025)	0.434 (0.025)
50	0%	0.329 (0.022)	0.330 (0.022)
	30%	0.491 (0.023)	0.493 (0.023)

Table 1.4: Explained variation of $Z|T$ versus $T|Z$; in () are standard errors from simulation runs.

β	1	3	15	50	100	1000
$R^2_{Z\|T}$	0.099 (0.020)	0.291 (0.024)	0.668 (0.026)	0.851 (0.020)	0.911 (0.017)	0.988 (0.006)
R^2_∞	0.090 (0.016)	0.208 (0.017)	0.308 (0.022)	0.328 (0.021)	0.332 (0.021)	0.333 (0.020)

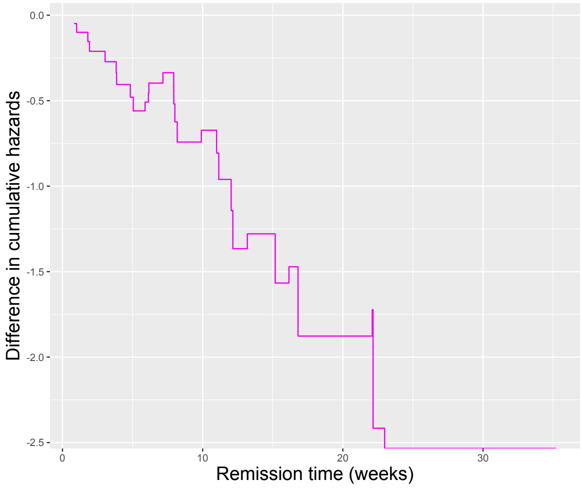


Figure 1.3: Difference between the cumulative hazard functions of the two groups for the Freireich data.

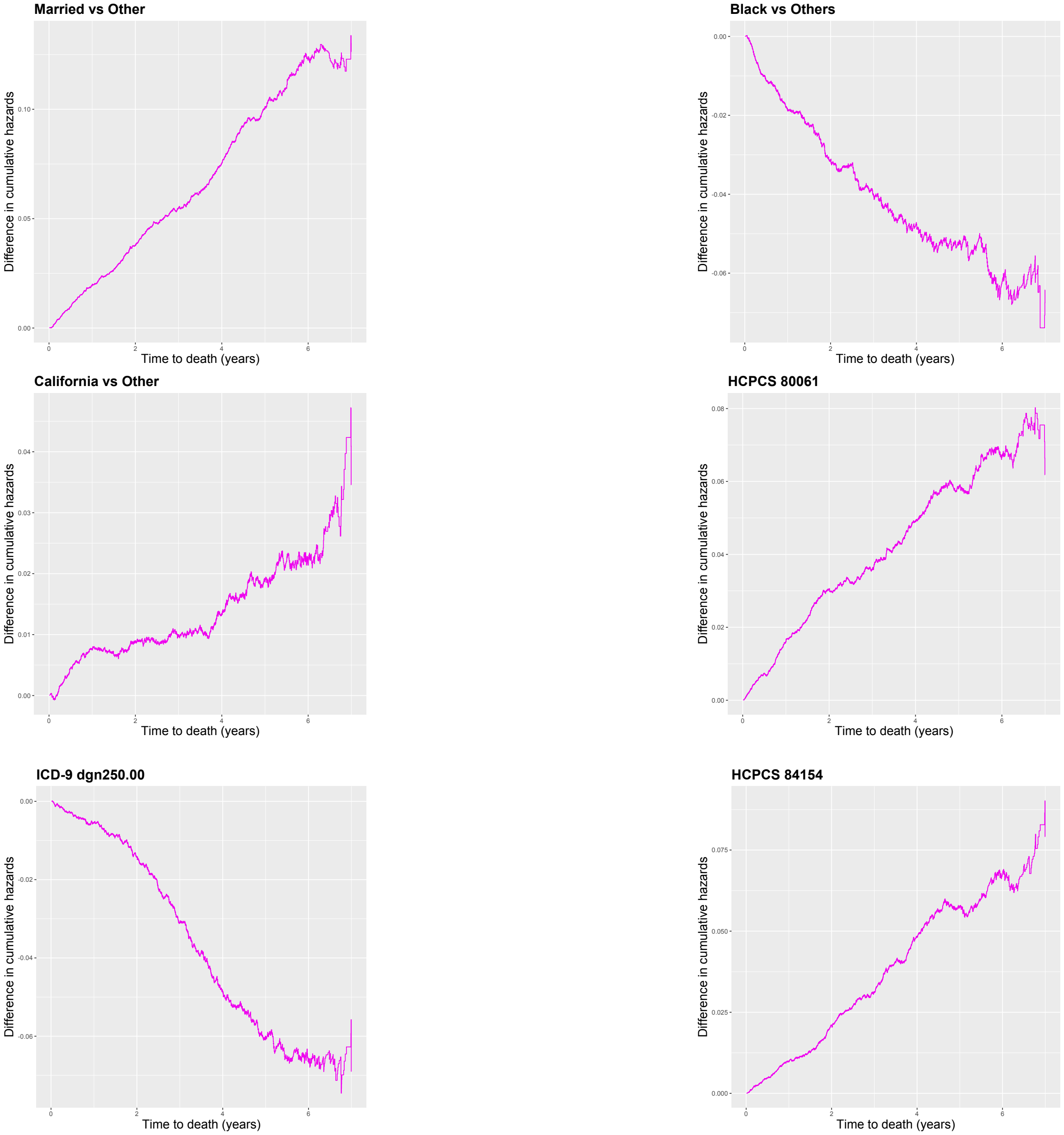


Figure 1.4: Difference between the cumulative hazard functions of groups defined by some dichotomous variables for the SEER-MEDICARE data.

Table 1.5: R^2 values for the SEER-Medicare data set. R^2 is computed on the full data set; R^2_{adj} is the adjusted R^2 also computed on the full data set; R^2_{out} is the out-of-sample R^2 computed on the test data set, with all parameters estimated from the training data set; and R^2_{train} is computed only on the training data set.

Model	R^2	R^2_{adj}	R^2_{out}	R^2_{train}
Clinical	0.048	0.048	0.053	0.051
Clinical + Demo.	0.270	0.270	0.271	0.261
Clinical + Demo. + Claims	0.373	0.370	0.388	0.379

1.8 Appendix

1.8.1 Ω^2 increases with $|\beta|$: proof of a specific case

Here we prove that Ω^2 increases with $|\beta|$ when Z is Bernoulii with $p = 0.5$ and under the semiparametric hazards model (1.1). We have:

$$\begin{aligned} E\{\mathrm{Var}(T \mid Z)\} &= E\left\{E\left(T^2 \mid Z\right)\right\} - \left[E\left\{E\left(T^2 \mid Z\right)\right\}\right]^2 \qquad (1.21) \\ &= \frac{1}{2}\left[2\int_0^\infty t\exp\{-\Lambda_0(t)\}dt - \left[\int_0^\infty \exp\{-\Lambda_0(t)\}dt\right]^2\right] \\ &\quad +\frac{1}{2}\left[2\int_0^\infty t\exp\{-\Lambda_0(t)-\beta t\}dt - \left[\int_0^\infty \exp\{-\Lambda_0(t)-\beta t\}dt\right]^2\right], \end{aligned}$$

and

$$
\begin{aligned}
\operatorname{Var}(T) &= E\left\{\operatorname{Var}\left(T \mid Z\right)\right\} + \operatorname{Var}\left\{E\left(T \mid Z\right)\right\} \qquad (1.22)\\
&= E\left\{E\left(T^2 \mid Z\right)\right\} - \left[E\left\{E\left(T \mid Z\right)\right\}\right]^2\\
&= \frac{1}{2}\left[2\int_0^\infty t\exp\left\{-\Lambda_0(t)\right\}dt + 2\int_0^\infty t\exp\left\{-\Lambda_0(t)-\beta t\right\}dt\right]\\
&\quad -\left[\frac{1}{2}\int_0^\infty \exp\left\{-\Lambda_0(t)\right\}dt + \frac{1}{2}\int_0^\infty \exp\left\{-\Lambda_0(t)-\beta t\right\}dt\right]^2.
\end{aligned}
$$

If we take the derivative with respect to $|\beta|$ of these quantities we get:

$$
\begin{aligned}
\frac{\partial E\left\{\operatorname{Var}\left(T \mid Z\right)\right\}}{\partial|\beta|} &= -\operatorname{sign}(\beta)\int_0^\infty t^2\exp\left\{-\Lambda_0(t)-\beta t\right\}dt \qquad (1.23)\\
&\quad +\operatorname{sign}(\beta)\int_0^\infty \exp\left\{-\Lambda_0(t)-\beta t\right\}dt\int_0^\infty t\exp\left\{-\Lambda_0(t)-\beta t\right\}dt\\
&= -\operatorname{sign}(\beta)\frac{1}{3}E\left(T^3 \mid Z=1\right) + \operatorname{sign}(\beta)\frac{1}{2}E\left(T \mid Z=1\right)E\left(T^2 \mid Z=1\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial \operatorname{Var}(T)}{\partial|\beta|} &= -\operatorname{sign}(\beta)\int_0^\infty t^2\exp\left\{-\Lambda_0(t)-\beta t\right\}dt \qquad (1.24)\\
&\quad +\operatorname{sign}(\beta)\left[\frac{1}{2}\int_0^\infty \exp\left\{-\Lambda_0(t)\right\}dt + \frac{1}{2}\int_0^\infty \exp\left\{-\Lambda_0(t)-\beta t\right\}dt\right]\\
&\quad \times\int_0^\infty t\exp\left\{-\Lambda_0(t)-\beta t\right\}dt\\
&= -\operatorname{sign}(\beta)\frac{1}{3}E\left(T^3 \mid Z=1\right) + \operatorname{sign}(\beta)\frac{1}{2}\left[\frac{1}{2}E\left(T \mid Z=1\right) + \frac{1}{2}E\left(T \mid Z=0\right)\right]\\
&\quad \times E\left(T^2 \mid Z=1\right).
\end{aligned}
$$

By equation (1.23) and (1.24), and after some algebra:

$$\begin{aligned}
& \frac{\partial \Omega_{\infty}^{2}}{\partial |\beta|} && (1.25)\\
= \; & -\text{sign}(\beta)\,(\text{Var}(T))^{-2} \\
& \times \left[E\left\{ \text{Var}\left(T \mid Z\right)\right\} \left\{ \frac{1}{4} E\left(T \mid Z=1\right) E\left(T^{2} \mid Z=1\right) - \frac{1}{4} E\left(T \mid Z=0\right) E\left(T^{2} \mid Z=1\right) \right\} \right] \\
& -\text{sign}(\beta) \frac{\text{Var}\left\{ E\left(T \mid Z\right)\right\} \left\{ -\frac{1}{3} E\left(T^{3} \mid Z=1\right) + \frac{1}{2} E\left(T \mid Z=1\right) E\left(T^{2} \mid Z=1\right)\right\}}{\left\{\text{Var}(T)\right\}^{2}}. && (1.26)
\end{aligned}$$

If now we consider the special case of $\lambda_0(t) = 1$, for which $\lambda(t) > 0$ if and only if $\beta > -1$, we have:

$$\begin{aligned}
& \frac{\partial \Omega_{\infty}^{2}}{\partial |\beta|} && (1.27)\\
= \; & \frac{\text{sign}(\beta)}{\left\{\text{Var}(T)\right\}^{2}} \left[\frac{1}{4} E\left\{ \text{Var}\left(T \mid Z\right)\right\} E\left(T^{2} \mid Z=1\right) \left(\frac{\beta}{1+\beta} \right) + \text{Var}\left\{ E\left(T \mid Z\right)\right\} \left(\frac{1}{(1+\beta)^{3}} \right) \right] \\
= \; & \frac{|\beta|\,(2+\beta)}{\left\{4\text{Var}(T)\right\}^{2}(1+\beta)^{4}} > 0, && (1.28)
\end{aligned}$$

proving that the measure increases with $|\beta|$.

1.8.2 SEER-MEDICARE Data

Table 1.6: Patient characteristics and claims codes from the SEER-Medicare dataset. Presented are mean (standard deviation) for the continuous variables, and frequency (%) for the binary variables.

	Overall ($n = 29657$)
PSA	11.34 (14.91)
GleasonScore	6.72 (0.94)
age	73.63 (5.61)
isBlack	3470 (11.7%)
isMarried	20501 (69.1%)
isRegCalifornia	13352 (45.0%)
var1001	162 (0.5%)
var7882	2026 (6.8%)
var5498	511 (1.7%)
var1806	2 (0.0%)
var17742	382 (1.3%)
var4270	24 (0.1%)
var4115	1 (0.0%)
var18195	12 (0.0%)
var4418	63 (0.2%)
var13233	29 (0.1%)
var4274	102 (0.3%)
var20250	53 (0.2%)
var4091	54 (0.2%)
var4074	25 (0.1%)
var4286	32 (0.1%)
var4137	2078 (7.0%)
var10944	226 (0.8%)
var4117	2902 (9.8%)
var3975	1169 (3.9%)
var4145	170 (0.6%)
var3890	322 (1.1%)
var4078	953 (3.2%)
var4229	816 (2.8%)
var4165	1364 (4.6%)
var4003	800 (2.7%)
var7750	1678 (5.7%)
var17042	3662 (12.3%)
var1517	18 (0.1%)
var1718	117 (0.4%)
var1456	837 (2.8%)
var1500	210 (0.7%)
var5681	2 (0.0%)
var14203	16 (0.1%)
var14388	31 (0.1%)
var5462	171 (0.6%)
var7887	431 (1.5%)
var5450	61 (0.2%)
var5525	1920 (6.5%)
var8019	24773 (83.5%)
var1921	76 (0.3%)
var13593	30 (0.1%)
var16169	99 (0.3%)
var1870	297 (1.0%)
var1844	3942 (13.3%)
var1361	45 (0.2%)
var15637	2 (0.0%)
var2485	6 (0.0%)
var17297	6 (0.0%)
var2431	43 (0.1%)
var2433	23 (0.1%)
var2426	357 (1.2%)
var7718	426 (1.4%)
var7673	356 (1.2%)
var3770	1 (0.0%)
var15060	1480 (5.0%)
var7684	2311 (7.8%)
var16918	1 (0.0%)
var3503	8 (0.0%)
var2833	496 (1.7%)
var17473	3 (0.0%)
var11745	6 (0.0%)
var6136	5 (0.0%)
var18556	4 (0.0%)
var16384	15 (0.1%)
var6164	2 (0.0%)
var4827	5 (0.0%)
var14591	12 (0.0%)
var21353	22 (0.1%)
var12800	4 (0.0%)
var18487	21 (0.1%)
var17644	21 (0.1%)
var10776	115 (0.4%)

	Overall ($n = 29657$)
var6937	28 (0.1%)
var15867	14 (0.0%)
var18793	17 (0.1%)
var19387	10 (0.0%)
var15736	65 (0.2%)
var11825	25 (0.1%)
var7025	37 (0.1%)
var18457	124 (0.4%)
var17613	91 (0.3%)
var7723	372 (1.3%)
var16068	601 (2.0%)
var19322	88 (0.3%)
var6170	276 (0.9%)
var17623	78 (0.3%)
var17574	1140 (3.8%)
var17743	548 (1.8%)
var11902	1648 (5.6%)
var17739	1985 (6.7%)
var15454	23500 (79.2%)
var17591	1022 (3.4%)
var16063	6742 (22.7%)
var17577	2680 (9.0%)
var19323	153 (0.5%)
var19342	97 (0.3%)
var21322	51 (0.2%)
var18854	2417 (8.1%)
var17734	10 (0.0%)
var1937	10 (0.0%)
var1927	6 (0.0%)
var1938	14 (0.0%)
var2082	5 (0.0%)
var1979	376 (1.3%)
var2100	223 (0.8%)
var4775	66 (0.2%)
var4706	3027 (10.2%)
var4663	696 (2.3%)
var4671	654 (2.2%)
var4769	386 (1.3%)
var17257	1101 (3.7%)
var4785	226 (0.8%)
var4758	419 (1.4%)
var9553	4 (0.0%)
var9945	5 (0.0%)
var10180	25 (0.1%)
var9082	7 (0.0%)
var10199	155 (0.5%)
var10902	1976 (6.7%)
var21288	16 (0.1%)
var5454	665 (2.2%)
var3873	2 (0.0%)
var4158	123 (0.4%)
var4339	1630 (5.5%)
var4282	702 (2.4%)
var1724	36 (0.1%)
var1455	6985 (23.6%)
var5456	391 (1.3%)
var5466	587 (2.0%)
var14419	24699 (83.3%)
var1810	3 (0.0%)
var840	5 (0.0%)
var2444	329 (1.1%)
var13506	3 (0.0%)
var21338	1 (0.0%)
var7810	36 (0.1%)
var16062	20290 (68.4%)
var15698	20071 (67.7%)

Table 1.7: Fit of the additive hazards model on the full data set with clinical variables as covariates

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
PSA	9.40e-04	5.83e-05	16.130	<0.001
GleasonScore	1.21e-02	7.13e-04	16.939	<0.001

Table 1.8: Fit of the additive hazards model on the full data set with clinical and demographical variables as covariates

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
PSA	7.13e-04	5.81e-05	12.266	<0.001
GleasonScore	8.94e-03	7.07e-04	12.657	<0.001
age	3.87e-03	1.34e-04	28.906	<0.001
isBlack	1.07e-02	1.83e-03	5.834	<0.001
isMarried	-1.21e-02	1.23e-03	-9.771	<0.001
isRegCalifornia	-4.05e-03	1.02e-03	-3.976	<0.001

Table 1.9: Fit of the additive hazards model on the full data set with clinical, demographical variables and claim codes as covariates.

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
PSA	5.65e-04	5.81e-05	9.725	<0.001
GleasonScore	8.34e-03	7.08e-04	11.775	<0.001
age	3.05e-03	1.34e-04	22.842	<0.001
isBlack	2.46e-03	1.85e-03	1.331	0.183
isMarried	-7.85e-03	1.23e-03	-6.353	<0.001
isRegCalifornia	-3.31e-03	1.05e-03	-3.145	0.002
var1001	3.47e-02	1.26e-02	2.749	0.006
var7882	7.67e-03	2.98e-03	2.570	0.010
var5498	7.69e-03	6.15e-03	1.250	0.211
var1806	-8.68e-03	9.52e-02	-0.091	0.927
var17742	3.01e-02	1.28e-02	2.346	0.019
var4270	-1.29e-02	3.34e-02	-0.385	0.700
var4115	3.47e-01	4.18e-01	0.829	0.407
var18195	1.38e-01	9.16e-02	1.508	0.132
var4418	1.06e-02	2.16e-02	0.492	0.623
var13233	6.05e-02	4.76e-02	1.270	0.204
var4274	2.50e-02	1.77e-02	1.409	0.159
var20250	6.05e-02	4.54e-02	1.332	0.183
var4091	1.16e-02	1.78e-02	0.653	0.514
var4074	1.52e-02	3.31e-02	0.459	0.646
var4286	2.09e-02	4.73e-02	0.442	0.658
var4137	2.98e-02	3.88e-03	7.671	<0.001
var10944	1.88e-02	1.23e-02	1.531	0.126
var4117	1.60e-02	2.58e-03	6.191	<0.001
var3975	5.70e-03	3.94e-03	1.446	0.148
var4145	3.59e-02	1.56e-02	2.294	0.022
var3890	3.27e-02	1.24e-02	2.637	0.008
var4078	8.66e-03	5.10e-03	1.697	0.090
var4229	1.27e-02	5.36e-03	2.370	0.018
var4165	3.94e-03	3.84e-03	1.027	0.305

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
var4003	7.87e-03	4.91e-03	1.602	0.109
var7750	3.60e-05	3.34e-03	0.011	0.991
var17042	-1.18e-02	1.51e-03	-7.783	<0.001
var1517	2.34e-02	3.54e-02	0.661	0.509
var1718	1.29e-02	2.23e-02	0.579	0.563
var1456	3.56e-03	4.53e-03	0.786	0.432
var1500	1.71e-03	1.18e-02	0.145	0.885
var5681	5.03e-02	1.03e-01	0.489	0.625
var14203	3.63e-02	4.17e-02	0.873	0.383
var14388	8.30e-02	4.08e-02	2.033	0.042
var5462	7.46e-03	1.61e-02	0.462	0.644
var7887	1.64e-02	7.53e-03	2.171	0.030
var5450	5.73e-02	3.38e-02	1.697	0.090
var5525	4.34e-03	3.28e-03	1.322	0.186
var8019	-4.36e-03	2.17e-03	-2.014	0.044
var1921	1.98e-02	1.69e-02	1.171	0.242
var13593	2.95e-02	3.36e-02	0.879	0.379
var16169	3.38e-02	1.77e-02	1.909	0.056
var1870	3.91e-03	8.24e-03	0.474	0.635
var1844	2.41e-03	1.98e-03	1.215	0.225
var1361	1.76e-02	2.28e-02	0.770	0.441
var15637	7.47e-01	6.39e-01	1.169	0.242
var2485	1.47e-01	9.78e-02	1.498	0.134
var17297	1.42e-01	1.02e-01	1.387	0.165
var2431	3.63e-02	2.62e-02	1.388	0.165
var2433	4.75e-02	4.95e-02	0.960	0.337
var2426	4.15e-02	1.16e-02	3.586	<0.001
var7718	-3.88e-03	6.76e-03	-0.573	0.567
var7673	2.45e-02	1.09e-02	2.259	0.024

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
var3770	-4.61e-02	3.85e-03	-11.990	<0.001
var15060	1.55e-03	4.23e-03	0.368	0.713
var7684	-8.48e-03	2.37e-03	-3.579	<0.001
var16918	3.02e-01	3.31e-01	0.915	0.360
var3503	7.69e-03	3.73e-02	0.206	0.837
var2833	6.05e-03	5.27e-03	1.148	0.251
var17473	1.21e-02	6.63e-02	0.183	0.855
var11745	9.91e-02	1.00e-01	0.989	0.322
var6136	2.68e-01	1.47e-01	1.826	0.068
var18556	3.24e-01	3.18e-01	1.020	0.308
var16384	2.71e-02	3.19e-02	0.848	0.396
var6164	7.30e-02	1.16e-01	0.628	0.530
var4827	6.75e-02	7.98e-02	0.846	0.397
var14591	2.58e-03	3.33e-02	0.078	0.938
var21353	2.53e-02	5.98e-02	0.424	0.672
var12800	1.77e-01	2.34e-01	0.756	0.450
var18487	5.82e-03	3.19e-02	0.183	0.855
var17644	9.72e-02	6.25e-02	1.555	0.120
var10776	2.19e-02	1.54e-02	1.426	0.154
var6937	1.72e-02	2.59e-02	0.663	0.507
var15867	2.48e-02	3.58e-02	0.694	0.488
var18793	1.93e-01	9.11e-02	2.119	0.034
var19387	2.64e-01	1.48e-01	1.778	0.075
var15736	3.85e-02	2.12e-02	1.813	0.070
var11825	1.04e-01	5.80e-02	1.801	0.072
var7025	2.03e-02	3.83e-02	0.529	0.597
var18457	2.76e-02	2.09e-02	1.320	0.187
var17613	1.20e-02	2.36e-02	0.510	0.610
var7723	7.24e-03	8.27e-03	0.875	0.381

	Estimate	Std. Error	Z value	$Pr(> \lvert z \rvert)$
var16068	-5.10e-03	5.03e-03	-1.013	0.311
var19322	-1.87e-02	2.68e-02	-0.700	0.484
var6170	8.18e-03	9.07e-03	0.902	0.367
var17623	-3.19e-02	2.67e-02	-1.192	0.233
var17574	-7.50e-04	4.94e-03	-0.152	0.879
var17743	2.03e-02	8.84e-03	2.293	0.022
var11902	6.65e-03	3.63e-03	1.833	0.067
var17739	2.61e-04	4.45e-03	0.059	0.953
var15454	-5.19e-03	2.90e-03	-1.788	0.074
var17591	-1.23e-03	5.75e-03	-0.214	0.831
var16063	-1.90e-03	1.16e-03	-1.638	0.101
var17577	-2.83e-03	3.46e-03	-0.818	0.413
var19323	3.51e-02	1.99e-02	1.761	0.078
var19342	4.84e-02	3.17e-02	1.528	0.126
var21322	-1.01e-02	3.55e-02	-0.284	0.777
var18854	-6.69e-03	1.73e-03	-3.855	<0.001
var17734	3.85e-02	5.37e-02	0.718	0.473
var1937	2.58e-01	1.91e-01	1.349	0.177
var1927	2.48e-01	1.74e-01	1.422	0.155
var1938	1.41e-01	1.08e-01	1.312	0.189
var2082	-7.31e-02	4.14e-02	-1.765	0.078
var1979	4.88e-02	1.26e-02	3.873	<0.001
var2100	1.02e-02	1.33e-02	0.768	0.442
var4775	6.64e-02	2.77e-02	2.399	0.016
var4706	1.32e-02	2.60e-03	5.087	<0.001
var4663	1.87e-02	6.19e-03	3.023	0.003
var4671	7.59e-03	5.85e-03	1.298	0.194
var4769	1.07e-02	7.42e-03	1.447	0.148
var17257	8.46e-03	3.78e-03	2.240	0.025

	Estimate	Std. Error	Z value	$Pr(> \lvert z \rvert)$
var4785	2.53e-02	1.23e-02	2.065	0.039
var4758	2.51e-02	9.74e-03	2.581	0.010
var9553	1.42e-01	9.64e-02	1.469	0.142
var9945	-1.35e-02	9.99e-02	-0.135	0.893
var10180	2.74e-02	2.58e-02	1.061	0.288
var9082	-1.09e-02	1.10e-01	-0.099	0.921
var10199	2.04e-02	1.27e-02	1.608	0.108
var10902	4.29e-03	2.68e-03	1.600	0.110
var21288	2.54e-02	5.33e-02	0.476	0.634
var5454	-1.60e-03	8.14e-03	-0.196	0.845
var3873	2.60e-01	2.25e-01	1.156	0.248
var4158	6.33e-03	1.64e-02	0.385	0.700
var4339	4.61e-03	3.45e-03	1.336	0.182
var4282	1.11e-02	6.04e-03	1.834	0.067
var1724	-3.30e-02	2.17e-02	-1.520	0.129
var1455	2.27e-03	1.41e-03	1.613	0.107
var5456	1.41e-02	7.90e-03	1.787	0.074
var5466	1.77e-02	7.63e-03	2.324	0.020
var14419	-2.02e-03	3.55e-03	-0.569	0.570
var1810	1.86e-01	2.19e-01	0.849	0.396
var840	2.83e-02	7.38e-02	0.384	0.701
var2444	3.65e-02	9.14e-03	3.990	<0.001
var13506	4.94e-01	3.49e-01	1.415	0.157
var21338	3.08e-01	3.73e-01	0.828	0.408
var7810	6.23e-02	4.80e-02	1.299	0.194
var16062	-1.77e-03	1.36e-03	-1.302	0.193
var15698	-8.28e-03	1.36e-03	-6.084	<0.001
var17681	-3.43e-03	1.68e-03	-2.042	0.041
var7826	1.13e-04	2.36e-03	0.048	0.962

Table 1.10: Fit of the additive hazards model on the training data set with clinical variables as covariates

	Estimate	Std. Error	Z value	$Pr(> \lvert z \rvert)$
PSA	8.97e-04	8.02e-05	11.173	<0.001
GleasonScore	1.38e-02	1.04e-03	13.191	<0.001

Table 1.11: Fit of the additive hazards model on the training data set with clinical and demographical variables as covariates

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
PSA	6.93e-04	8.01e-05	8.650	<0.001
GleasonScore	1.04e-02	1.03e-03	10.060	<0.001
age	3.87e-03	1.90e-04	20.356	<0.001
isBlack	9.14e-03	2.60e-03	3.511	<0.001
isMarried	-1.17e-02	1.75e-03	-6.698	<0.001
isRegCalifornia	-3.41e-03	1.47e-03	-2.329	0.020

Table 1.12: Fit of the additive hazards model on the training data set with clinical, demographical variables and claim codes as covariates

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
PSA	5.43e-04	7.99e-05	6.795	<0.001
GleasonScore	9.92e-03	1.03e-03	9.594	<0.001
age	3.04e-03	1.91e-04	15.936	<0.001
isBlack	1.65e-03	2.63e-03	0.627	0.530
isMarried	-7.85e-03	1.75e-03	-4.487	<0.001
isRegCalifornia	-3.29e-03	1.52e-03	-2.166	0.030
var1001	3.33e-02	1.80e-02	1.848	0.065
var7882	6.30e-03	4.22e-03	1.492	0.136
var5498	1.41e-02	9.21e-03	1.527	0.127
var1806	1.18e-01	2.24e-01	0.529	0.597
var17742	3.91e-02	1.84e-02	2.128	0.033
var4270	-4.00e-02	3.61e-02	-1.109	0.267
var4115	3.37e-01	4.18e-01	0.806	0.420
var18195	1.95e-02	9.91e-02	0.197	0.844
var4418	1.09e-02	2.53e-02	0.429	0.668
var13233	9.14e-02	6.35e-02	1.439	0.150
var4274	2.41e-02	2.22e-02	1.089	0.276
var20250	2.58e-02	5.94e-02	0.434	0.664
var4091	4.92e-02	3.54e-02	1.387	0.165
var4074	-8.89e-03	3.98e-02	-0.223	0.823
var4286	3.52e-02	7.59e-02	0.463	0.643
var4137	3.28e-02	5.60e-03	5.849	<0.001
var10944	7.20e-03	1.68e-02	0.427	0.669
var4117	2.17e-02	3.76e-03	5.767	<0.001
var3975	9.11e-03	5.66e-03	1.609	0.108
var4145	5.65e-02	2.52e-02	2.245	0.025
var3890	3.03e-02	1.79e-02	1.695	0.090
var4078	1.17e-02	7.85e-03	1.492	0.136
var4229	1.11e-02	7.41e-03	1.495	0.135
var4165	3.78e-04	5.54e-03	0.068	0.946

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
var4003	5.79e-03	6.80e-03	0.852	0.394
var7750	-8.45e-04	4.72e-03	-0.179	0.858
var17042	-1.18e-02	2.21e-03	-5.337	<0.001
var1517	3.72e-02	4.82e-02	0.771	0.441
var1718	1.97e-02	3.23e-02	0.609	0.543
var1456	-4.07e-03	6.01e-03	-0.677	0.499
var1500	-1.18e-03	1.75e-02	-0.068	0.946
var5681	4.29e-02	1.03e-01	0.419	0.675
var14203	1.12e-01	9.42e-02	1.193	0.233
var14388	2.05e-01	9.74e-02	2.101	0.036
var5462	1.85e-02	2.30e-02	0.807	0.420
var7887	2.25e-02	1.07e-02	2.099	0.036
var5450	8.41e-02	5.89e-02	1.427	0.154
var5525	1.00e-03	4.54e-03	0.221	0.825
var8019	-4.03e-03	3.15e-03	-1.281	0.200
var1921	3.55e-02	2.30e-02	1.545	0.122
var13593	2.98e-02	4.32e-02	0.691	0.490
var16169	7.19e-02	3.19e-02	2.254	0.024
var1870	1.06e-02	1.27e-02	0.834	0.404
var1844	1.59e-03	2.82e-03	0.566	0.572
var1361	-1.94e-02	2.78e-02	-0.699	0.485
var15637	7.79e-01	6.41e-01	1.215	0.224
var2485	2.76e-01	1.86e-01	1.488	0.137
var17297	2.27e-01	1.56e-01	1.458	0.145
var2431	1.95e-02	3.17e-02	0.616	0.538
var2433	-1.50e-03	5.20e-02	-0.029	0.977
var2426	4.27e-02	1.53e-02	2.786	0.005
var7718	-2.70e-03	9.41e-03	-0.287	0.774
var7673	4.05e-02	1.64e-02	2.471	0.013

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
var3770	-4.34e-02	5.46e-03	-7.956	<0.001
var15060	-7.18e-04	5.77e-03	-0.124	0.901
var7684	-7.28e-03	3.36e-03	-2.165	0.030
var16918	3.03e-01	3.31e-01	0.916	0.360
var3503	-3.27e-02	7.41e-03	-4.412	<0.001
var2833	7.53e-03	7.32e-03	1.028	0.304
var17473	-1.78e-02	5.24e-03	-3.394	0.001
var11745	4.99e-02	9.56e-02	0.521	0.602
var6136	4.32e-01	2.30e-01	1.878	0.060
var18556	5.42e-02	2.54e-01	0.213	0.831
var16384	1.70e-02	4.81e-02	0.354	0.724
var6164	2.49e-01	3.33e-01	0.749	0.454
var4827	6.43e-02	8.00e-02	0.803	0.422
var14591	-6.18e-03	4.02e-02	-0.154	0.878
var21353	-4.77e-03	8.93e-02	-0.053	0.957
var12800	2.28e-01	2.90e-01	0.786	0.432
var18487	2.37e-03	3.88e-02	0.061	0.951
var17644	1.45e-01	9.12e-02	1.591	0.112
var10776	1.00e-02	1.91e-02	0.525	0.600
var6937	4.35e-02	5.22e-02	0.833	0.405
var15867	2.34e-02	6.42e-02	0.365	0.715
var18793	1.96e-01	2.06e-01	0.951	0.342
var19387	7.10e-01	5.05e-01	1.404	0.160
var15736	3.28e-02	2.70e-02	1.215	0.224
var11825	6.01e-02	6.59e-02	0.913	0.361
var7025	4.87e-02	6.63e-02	0.735	0.463
var18457	5.25e-02	3.01e-02	1.745	0.081
var17613	-7.14e-03	3.28e-02	-0.218	0.828
var7723	5.64e-03	1.18e-02	0.478	0.633

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
var16068	-8.28e-03	6.90e-03	-1.199	0.231
var19322	-9.97e-03	4.07e-02	-0.245	0.807
var6170	-4.33e-04	1.20e-02	-0.036	0.971
var17623	-3.80e-02	3.73e-02	-1.018	0.309
var17574	-6.01e-03	6.94e-03	-0.866	0.387
var17743	2.11e-02	1.25e-02	1.686	0.092
var11902	3.46e-03	4.88e-03	0.709	0.478
var17739	-7.28e-03	6.14e-03	-1.186	0.236
var15454	-6.04e-03	4.29e-03	-1.408	0.159
var17591	-6.97e-03	8.33e-03	-0.838	0.402
var16063	-1.44e-03	1.68e-03	-0.859	0.391
var17577	-3.16e-04	5.01e-03	-0.063	0.950
var19323	1.48e-02	2.56e-02	0.577	0.564
var19342	5.97e-02	4.68e-02	1.278	0.201
var21322	3.09e-02	4.81e-02	0.642	0.521
var18854	-6.63e-03	2.45e-03	-2.705	0.007
var17734	4.81e-03	6.35e-02	0.076	0.940
var1937	2.34e+00	1.86e+00	1.256	0.209
var1927	2.27e-01	1.79e-01	1.263	0.206
var1938	1.16e-01	1.33e-01	0.868	0.385
var2082	-1.04e-01	6.37e-02	-1.631	0.103
var1979	3.62e-02	1.67e-02	2.162	0.031
var2100	9.01e-03	1.82e-02	0.494	0.621
var4775	6.17e-02	3.48e-02	1.775	0.076
var4706	1.27e-02	3.75e-03	3.395	0.001
var4663	2.12e-02	9.02e-03	2.355	0.019
var4671	-3.26e-03	8.08e-03	-0.403	0.687
var4769	2.84e-02	1.19e-02	2.382	0.017
var17257	7.48e-03	5.30e-03	1.412	0.158

	Estimate	Std. Error	Z value	$Pr(>\|z\|)$
var4785	1.69e-02	1.69e-02	1.001	0.317
var4758	4.04e-02	1.54e-02	2.617	0.009
var9553	-9.24e-03	8.65e-03	-1.068	0.285
var9945	-9.69e-02	8.91e-02	-1.088	0.277
var10180	3.36e-04	2.48e-02	0.014	0.989
var9082	1.41e-01	3.19e-01	0.443	0.658
var10199	2.77e-02	1.79e-02	1.550	0.121
var10902	-4.23e-04	3.75e-03	-0.113	0.910
var21288	-4.98e-02	6.76e-02	-0.737	0.461
var5454	1.30e-02	1.20e-02	1.085	0.278
var3873	2.62e-01	2.25e-01	1.166	0.244
var4158	4.49e-02	2.70e-02	1.663	0.096
var4339	3.74e-03	4.79e-03	0.780	0.435
var4282	1.57e-02	8.60e-03	1.824	0.068
var1724	-1.32e-02	3.27e-02	-0.402	0.687
var1455	4.35e-03	2.05e-03	2.120	0.034
var5456	9.20e-03	1.07e-02	0.862	0.389
var5466	1.63e-02	1.02e-02	1.595	0.111
var14419	-2.49e-03	5.30e-03	-0.470	0.639
var1810	7.76e-01	6.64e-01	1.168	0.243
var840	-4.77e-02	8.45e-03	-5.643	<0.001
var2444	4.58e-02	1.38e-02	3.311	0.001
var13506	5.10e-01	5.36e-01	0.952	0.341
var21338	3.12e-01	3.73e-01	0.838	0.402
var7810	5.60e-02	1.01e-01	0.552	0.581
var16062	-2.19e-04	1.90e-03	-0.115	0.908
var15698	-1.08e-02	1.96e-03	-5.523	<0.001
var17681	-2.68e-03	2.54e-03	-1.056	0.291
var7826	-4.38e-04	3.34e-03	-0.131	0.896

1.9 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in Statistics in Medicine. Rava, Denise; Xu, Ronghui. Explained variation under the additive hazards model, 40.1:85-100,2021. The dissertation author was the primary investigator and author of this paper.

Chapter 2

DeepHazard: neural network for time-varying risks

2.1 Introduction

Understanding the relationship between covariates and the distribution of survival time is fundamental in many fields spanning medicine, biology, healthcare, economics, and engineering. Survival data are often incomplete due to censoring, making the traditional predictive methods unsuitable. Traditionally, several semiparametric survival models, such as the popular Cox Model (Cox, 1972), the Additive Hazards Model (Aalen, 1980) or the Accelerated Failure Time model (Wei, 1992), have been proposed and extensively used. Developed to deal with censoring; however, they model the hazards as a particular function of a linear combination of the data, limiting their applicability in many real-world applications.

To overcome this difficulty, the interest in using deep learning methods, such as neural networks, for survival prediction has been increasing. Several nonparametric extensions of the Cox

Model have appeared in the literature; see, for example, (Faraggi and Simon, 1995; Ching et al., 2018; Liao and Ahn, 2016; Zhu et al., 2016; Katzman et al., 2018; Kvamme et al., 2019). They make use, to train the neural network, of the classical Cox partial likelihood and base their analysis on the proportional hazard assumption. The latter is often unrealistic and represents a relevant limitation. Non-proportional hazards are widely occurrent: when the effect of a treatment vanishes over time, and henceforth the ratio of the hazards tends to one, or when a drug is beneficial for one subgroup but harmful for the other, resulting in crossing survival curves. Non-proportional hazards are difficult to model. They usually indeed don't allow the use of a flexible and nonparametric baseline hazard.

Another line of work pertains the usage of discrete-time hazards for survival prediction; see for example, Liestbl et al. (1994); Brown et al. (1997); Biganzoli et al. (1998); Zhu et al. (2016); Luck et al. (2017); Fotso (2018); Lee et al. (2018); Gensheimer and Narasimhan (2019); Grisan et al. (2019); Ren et al. (2019); Zhao and Feng (2019); Lee et al. (2019). They don't make assumptions on the form of the hazard; however, they treat survival time as a discrete random variable taking only finitely many pre-determined values, loosing, therefore, the continuous nature of the problem itself. Moreover, they often cast the survival problem as a classification one, considering every observation as a sequence of zeros and ones to indicate their status. Naturally, with discrete approaches, the hazard is no longer a rate but a conditional probability. A different approach is the one proposed by Zhao and Feng (2019). The authors reduce the survival problem to a standard regression problem by considering inputting the missing outcomes with Kaplan Meier survival estimates. However, regression on such pseudo-responses is deemed biased whenever data is not missing at random. We construct, instead, a new survival neural network.

To overcome these limitations, we propose DeepHazard, a new neural network that doesn't rely on the assumption of proportional hazards while not neglecting the continuous nature of the data. Our approach is indeed tailored for a wide range of hazards, with the only restriction of being continuous and additive in time. Illustrative examples include a case where the effect of treatment or the treatment status changes with time; some patients are treated only after their disease progresses.

Building on the promising alternative of the Cox model, the non-parametric additive hazards model, we propose a new non-parametric alternative of the additive hazards loss. The latter doesn't constrain the risk of being of a particular form or being constant in time. Moreover, it naturally incorporates time-dependent covariates making our approach suitable for a large class of real data applications. In particular, our approach is designed to treat an aligned type of data arising whenever for each observation, and each covariate, a sequence of measurement at different time points is available, for example, in a series of follow-up visits.

The sequential nature of the data is incorporated by dividing the data in multiple time-frames and building a neural network in each time-frame to estimate the time-varying risk. Each neural network is trained on the observations, still at risk. Moreover, the interdependency between different time-frames is directly assimilated by adding to the input of every time interval-specific neural network, the output of the network built in the previous time-period. Figure 2.1 presents one possible architecture. The output node (blue in Figure 2.1) of each of the time-frames, denotes the predicted value of the risk score at that period. The input nodes (red in Figure 2.1) in each of the time-frames denote at-risk observations at that time-frame. Note that they change both in numbers and type from time-frame to time-frame. For the proposed neural network, the steps of feature extraction and survival analysis are not separated or done through two separate optimization procedures. They are gathered in one unique neural network, and the optimization of all the parameters happens together using the proposed survival loss. In this way, observations still at-risk are kept together.

DeepHazard outputs, for each combination of covariates, a rich estimate of the risk function and, for external covariates, survival function, including the baseline survival, as well as survival in desired time-intervals, therefore allowing a deep understanding of the time to event distribution and comparison between different groups and observations. The performance of our approach is evaluated through extensive simulations. We show that our method outperforms existing methods in terms of predictive capability, evaluated through the time-dependent C-index metric (Antolini et al., 2005). We also apply DeepHazard to the popular real datasets: METABRIC, GBSG, and ACTG to study time to death of breast-cancer and HIV-infected patients.

2.1.1 Related literature

Different methods that make use of machine learning techniques have been employed to analyze continuous survival data. Random survival forest of Ishwaran et al. (2008) extends the random forest methodology to survival analysis. Recently broadened to accommodate time-varying covariates, (Wongvibulsin et al., 2020), random survival forest consists of an ensemble of survival trees that are grown following a particular splitting rule that aims to maximize the difference between estimated survival curves in children nodes. Although a model is not explicitly assumed, the random survival forest's predictive performance depends on the splitting rule chosen. The most popular uses log-rank split statistics, which is known to lack power when the proportional hazards assumption is violated.

Machine learning techniques for discrete-time survival data include DeepHit and Dynamic-DeepHit, (Lee et al., 2018, 2019), a neural network that directly estimates the probability mass function of experiencing a particular event at a specific time. Fotso (2018) recasts the output of observation as a sequence of zeros (up to the event time) followed by a sequence of ones (after the event time) and applies the framework of neural networks to the multi-task logistic regression. Kvamme and Borgan (2019) rewrites the output as a vector of zeros with a single one corresponding to the observed event and makes use of the negative log-likelihood for Bernoulli data to train the neural network. The authors then propose an extension to continuous-time survival data using discretization and interpolation strategies. Zhong and Tibshirani (2019) introduce the stacking idea that recasts the data into a large data frame where the output column is a series of zeros and ones. The problem is then treated as a classification problem onto which various existing techniques can be directly applied.

When the time is not discretized and is treated as continuous, semi-parametric approaches based on the popular Cox model have been proposed. Katzman et al. (2018) parametrizes a Cox regression model with a neural network building on the work of Faraggi and Simon (1995). Kvamme et al. (2019) proposes an extension of it introducing an approximation of the partial log-likelihood to batches of data and allowing the relative risk function to depend on time. In both cases, the model is a relative risk model that does not allow the introduction of time-dependent covariates. A fully parametric approach has recently been proposed by Nagpal et al. (2020), where the survival function conditional on the fixed (not time-dependent) covariates is assumed to be a mixture of individual parametric survival distributions.

In this work, we build on the literature of semiparametric models for continuous-time survival data, proposing a different loss function, entirely unrelated to the partial likelihood typical of the proportional hazards model. Moreover, we propose a framework that allows the extension of our and potentially many other neural network methodologies to time-dependent covariates.

2.1.2 Organization of the paper

Section 2.2 contains the details of the proposed DeepHazard algorithm which includes a new time-additive hazards model, Section 2.2.1, a decomposition of the loss function, Section 2.2.3, as well as the details of the estimation and prediction, Section 2.2.4 and 2.2.5, respectively. Section 2.3 includes detailed finite sample experiments on time-dependent covariates and outcomes where we illustrate the impact of censoring, sample size, time, and feature space. Section 2.4 focuses on real data examples where we compare with the Random Survival Forest and DeepSurv algorithms and demonstrate superior performance.

2.2 DeepHazard learning

We introduce a new survival model, additive in time only, that explains the survival of a subject given, possibly time-varying, covariates.

Observations of survival times are often censored. This is the case when a patient drops out of a hospital or drug-treatment study. The time of death is, in this case, never observed; however, we know that the patient was still alive when he left the study. This is modeled with a random variable C. If T denotes the survival time, then the censored observations regarding the outcome of interest are denoted with $X = \min\{T, C\}$. Together with X we typically assume that an event indicator, δ is observed; here, $\delta = \mathbb{1}\{T \leq C\}$.

Medical studies are typically monitored in regular time intervals where a set of personal, medical information is collected, such as blood pressure, drugs taken, temperature reading, oxygenation of the blood. Some of those can naturally be treated as baseline variables, i.e., variables not changing with time; examples include gene expressions of particular tumor tissue, demographics, age. However, the majority are time-varying. For simplicity in notation, we denote all of the covariates as time-varying variables $Z(t) \in \mathbb{R}^p$.

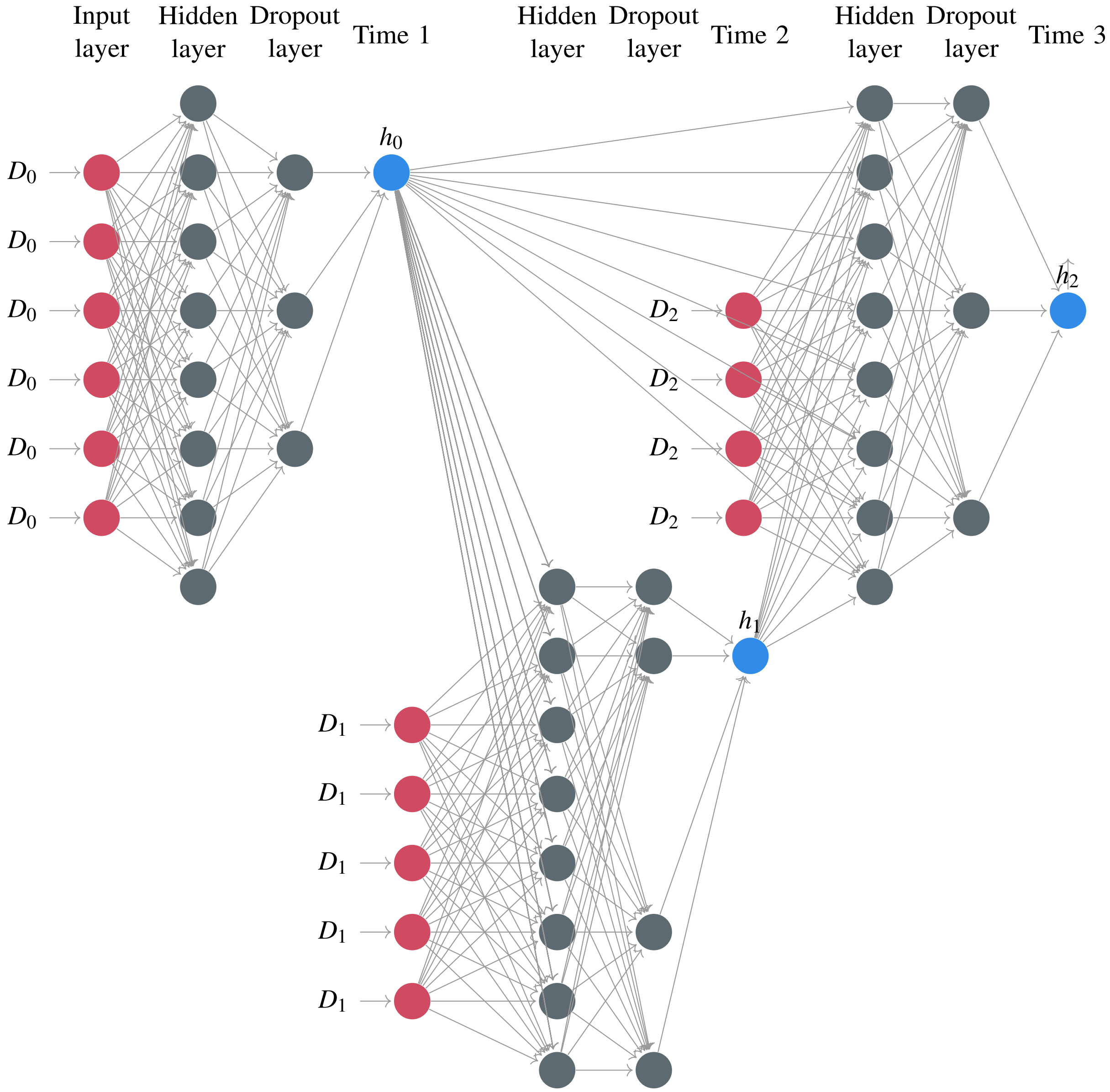


Figure 2.1: Example of DeepHazard architecture: The output node (blue) of each time-frame, denotes the predicted value of the hazard at that time period. The input nodes (red) in each of the time-frames denote at-risk observations at that time-frame.

2.2.1 Time additive hazards model

We propose a new model, additive in time, that assumes that the hazard function,

$$\lambda(t \mid Z(t)) = \lim_{h \to 0} \frac{P\big(T \in (t, t+h] \big| T \geq t, Z(t)\big)}{h},$$

is the sum of two components, a baseline hazard $\lambda_0(t)$ that depends only on time and a risk score, $h(Z(t),t)$ that encloses the effect of the individual's covariates $Z(t)$, possibly time-varying, onto the hazard. The hazard is interpreted in the standard way, as the probability of an event in the interval $[t, t+dt)$ given covariate $Z(t)$ and assuming that no previous event has happened.

We assume that the covariates are measured at a sequence of M time points (follow up visits),

$$t_0, t_1, \ldots, t_M.$$

Let's notice that we don't require $t_0, t_1, \ldots, t_M$ to be the same as event times. Therefore, we naturally divide the time into a sequence of intervals $[t_0, t_1), \ldots, [t_M, \infty)$. For example, let us assume that every patient is subjected to a visit every two months, and at every such visit, a series of physical values, as blood pressure, is measured and recorded. In this case, we would have as intervals $[0,2), [2,4), \ldots$, and the series of the measured values will be encoded as $Z(0), Z(2), \ldots$. We assume that at any intervals $[t_j, t_{j+1})$ the risk score of a subject is described by a constant in time risk score h_j.

$$h(Z(t),t) = h_j(Z(t)), \qquad t \in [t_j, t_{j+1}), \qquad j = 0, 1, \ldots, M,$$

To acknowledge the continuous nature of the time and the natural possible dependence onto the past values, we allow the risk score h_j to depend on previous-in-time risk scores $h_0, \ldots, h_{j-1}$. In other words $h(Z(t),t)$ satisfies

$$h_j(Z(t)) = f_j(Z(t_j), h_0(Z(t)), h_1(Z(t)), \cdots, h_{j-1}(Z(t))), \quad t \in [t_j, t_{j+1}), \quad j = 0, 1, \ldots, M, \quad (2.1)$$

where f_j is an unknown function. This describes a recursive relationship

$$\begin{aligned} &h_0(Z(t)) = f_0(Z(t_0)),\ h_1(Z(t)) = f_1\big(Z(t_1), f_0(Z(t_0))\big), \\ &h_2(Z(t)) = f_2\Big(Z(t_2), f_0(Z(t_0)), f_1\Big(Z(t_1), f_0(Z(t_0))\Big)\Big), \cdots. \end{aligned}$$

With a small abuse in notation we drop the notation f_j and use h_j to denote the unknown functional relationship at time interval j.

Therefore, primarily we consider the following representation of the hazard

$$\lambda(t|Z(t)) = \lambda_0(t) + h(Z(t),t) \quad (2.2)$$

where

$$h(Z(t),t) = \sum_{j=0}^{M} h_j\Big(Z(t_j), h_0(Z(t)), \ldots, h_{j-1}(Z(t))\Big) \mathbb{1}\left(t_j \le t < t_{j+1}\right), \quad (2.3)$$

where $t_{M+1} = \infty$ and $h_0(Z(t)), \ldots, h_M(Z(t))$ are functions of the covariates.

The form of model (3.1) is reminiscent of the traditional additive hazards model (Aalen, 1980), which takes the following form, $\lambda(t \mid Z) = \lambda_0(t) + \beta(t)Z(t)$ with the risk being limited to be of a linear form. The proposed model extends it to comprise a broader range of risk score forms and to incorporate the sequential nature of time-varying covariates.

Example 1: Sum of all previous in time hazards:

$$h_0(Z(t)) = f_0(Z(t_0)),\ h_1(Z(t)) = h_0(Z(t)) + f_1(Z(t_1)),$$
$$h_j(Z(t)) = h_0(Z(t)) + \cdots + h_{j-1}(Z(t)) + f_j(Z(t_j))$$

This can be named nonparametric additive hazards model; structure of the hazard mimics that of generalized additive models (Hastie and Tibshirani, 1990). Similarly, one can consider sum of the last few in time hazards only.

Example 2: Product of the last k hazards:

$$h_0(Z(t)) = f_0(Z(t_0)),\ h_1(Z(t)) = f_0(Z(t_0))f_1(Z(t_1)),$$
$$h_j(Z(t)) = h_{j-k}(Z(t)) \cdots h_{j-1}(Z(t)) f_j(Z(t_j)).$$

Here the logarithm of the hazard has nonparametric and additive structure. However the logarithmic transformation as well as functions $f_0, \ldots, f_M$ are unknown a-priori.

Example 3: Heterogeneous hazard:

$$h_j(Z(t)) = \sigma_j Z(t_j) \qquad \sigma_j^2 = \omega_j + \alpha_j f_{j-1}^2(Z(t_{j-1})) + \beta_j \sigma_{j-1}^2,$$

where $\omega > 0$, $\alpha, \beta \geq 0$. The aforementioned constants as well as functions f_j are all unknown parameters of the hazard. In particular,

$$h_0(Z(t)) = \sigma_0 Z(t_0), \sigma_0^2 = \omega_0 > 0,$$
$$h_1(Z(t)) = \sigma_1 Z(t_1), \sigma_1^2 = \omega_1 + \alpha_1 (f_0(Z(t_0)) - \theta_1 \omega_0)^2 + \beta \omega_0^2, \cdots$$

More in general, it is easy to see how any survival model can be written as equation (3.1). Our modeling Assumption, (2.3), on the form of $h(Z(t),t)$ can be seen as an approximation for estimation purposes. Indeed, we only assume the risk to be constant into intervals. Moreover, we allow the dependency of the risk score, of a specific interval, onto the risk scores of the previous intervals, making the assumption of piecewise constant risk less strict and allowing the continuous nature of the time to play an explicit role. Therefore, our model can be applied to a wide variety of risk score forms. As long as the intervals are dense enough, and the smoothness of the risk score is adequate, our approximation will work well.

2.2.2 Quadratic loss function

In this section we want to motivate our score function or loss function through a population perspective first. In the following we use $Y(t) = \mathbb{1}(X \geq t)$ to denote the at-risk indicator, i.e. subset of observations which are at time t still at risk of experiencing an "event," i.e., death. In addition, we indicate with $N(t) = \mathbb{1}(X \leq t, \delta = 1)$ the counting process of whether and when an "event" has occurred.

The estimation strategy borrows techniques from the additive hazards model and its least squares loss therefore landing itself particularly useful for neural-network approaches. If indeed, we consider the generic representation of the model (3.1),

$$dN(t) = \lambda(t \mid Z(t))Y(t)dt + dM(t) \tag{2.4}$$

where $M(t)$ is the associated martingale process, the following least squares loss, also called in the literature least-squares contrast, (Reynaud-Bouret et al., 2006), for a generic function $f(Z(t),t)$, can be easily derived:

$$\gamma(f) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f(Z_i(t),t)dN_i(t) + \frac{1}{2n}\sum_{i=1}^{n}\int_0^\tau f(Z_i(t),t)^2 Y_i(t)dt.$$

In the above, τ is an upper bound of time due to administrative censoring. Taking the expected value on both sides of (2.4) and considering the martingale decomposition, we get:

$$\mathrm{E}\{\gamma(f)\} = -\mathrm{E}\left\{\int_0^\tau f(Z_i(t),t)\lambda(t \mid Z_i(t))Y_i(t)dt\right\} + \frac{1}{2}\mathrm{E}\left\{\int_0^\tau f(Z_i(t),t)^2 Y_i(t)dt\right\}.$$

Defining with $\|\cdot\|_\mu$, the following norm: $\|g\|_\mu = \mathrm{E}\left\{\int_0^\tau f^2(Z(t),t)Y(t)dt\right\}$, we are left with

$$\begin{aligned}2\mathrm{E}\{\gamma(f)\} &= \mathrm{E}\left[\int_0^\tau \{f(Z(t),t) - \lambda(t \mid Z(t))\}^2 \lambda(t \mid Z(t))Y_i(t)dt\right] - \mathrm{E}\left\{\int_0^\tau \lambda(t \mid Z(t))^2 Y_i(t)dt\right\}\\ &= \|f(Z(t),t) - \lambda(t \mid Z(t))\|_\mu - \|\lambda(t \mid Z(t))\|_\mu .\end{aligned}$$

The latter justifies the minimization of the least squares contrast as estimation strategy for the hazard function $\lambda(t \mid Z(t))$, as explained in Comte et al. (2011). If we consider our additive form of the hazard (3.1), $f(Z(t),t) = \lambda_0(t) + h(Z(t),t)$, the loss can be decomposed as follows:

$$\gamma(f) = \gamma_1(\lambda_0) + \gamma_2(h) + \gamma_3(\lambda_0,h),$$

where

$$\gamma_1(\lambda_0) = \frac{1}{2n}\sum_{i=1}^{n}\int_0^\tau \left\{\lambda_0(t) + \bar{h}(t)\right\}^2 Y_i(t)dt - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\lambda_0(t) + \bar{h}(t)\right\} dN_i(t),$$

$$\gamma_2(h) = \frac{1}{2n}\sum_{i=1}^{n}\int_0^\tau \left\{h(Z_i(t),t) - \bar{h}(t)\right\}^2 Y_i(t)dt - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{h(Z_i(t),t) - \bar{h}(t)\right\} dN_i(t),$$

$$\gamma_3(\lambda_0,h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{h(Z_i(t),t) - \bar{h}(t)\right\}\left\{\lambda_0(t) + \bar{h}(t)\right\} Y_i(t)dt,$$

where

$$\bar{h}(t) = \frac{\sum_{i=1}^{n} h(Z_i(t),t)Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

By easy computation it can be proven that $\gamma_3(\lambda_0, h) = 0$. It is therefore suitable, for estimation of the risk $h(Z(t),t)$, to consider the minimization of $\gamma_2(h)$ solely. Details of the above decompositions can be found in Gaiffas et al. (2012). In our approach, we make use of a regularized version of $\gamma_2(h)$,

$$\min_{h} \{\gamma_2(h) + P(h)\},$$

where P is an appropriate penalty function of practitioners choice. We show the details in the next section.

2.2.3 Loss function decomposition

Noticing that time-dependent covariates are observed in a natural, sequential ordering, $t_0 \leq t_1 \leq \cdots \leq t_M$, and because of the assumed form of the hazard (2.3), to estimate the risk $h(Z(t),t)$, we need to estimate the various time-intervals specific risk h_j for $j = 0, \ldots, M$. Intuitively, it makes sense to involve in the estimation of each h_j, only the observations at risk on the jth interval, discarding everyone that is censored or have experienced the events before the start of that particular interval. In the following, we explain the mathematical arguments in detail.

In our approach, every h_j will be estimated by a neural network j, whose parameters, biases and weights, will be indexed by θ_j. In the following we use the generic θ to indicate the collection of $(\theta_0, \ldots, \theta_M)$ and we use h_θ to denote the dependency, explained in details later, of the final estimate of $h(Z(t),t)$ onto the parameters of the networks. Henceforth, we make use of the following regularized version of $\gamma_2(h)$:

$$\gamma_2(h_\theta) + \lambda \sum_{j=0}^{M} \left\|\theta_j\right\|_p, \tag{2.5}$$

where we implemented two norms: $p = 1,2$ to allow for both the Lasso and the Ridge penalty.

We observe that the integrals in (2.5) can be broken down as sums of $M+1$ integrals, one for each time intervals introduced above, as in the following:

$$\sum_{j=0}^{M} \mathcal{L}_j(\theta_j) + \lambda \left\|\theta_j\right\|_p, \tag{2.6}$$

where

$$\mathcal{L}_j(\theta_j) = (2n)^{-1} \sum_{i=1}^{n} \int_{t_j}^{t_{j+1}} \left(Y_i^j(t) \left[h_{\theta_j}(Z_i(t),t) - \bar{h}_{\theta_j}(t)\right]^2 dt - 2\left[h_{\theta_j}(Z_i(t),t) - \bar{h}_{\theta_j}(t)\right] dN_i^j(t), \right), \tag{2.7}$$

where

$$N_i^j(t) = \mathbb{1}(X_i \leq t, \delta_i = 1, t_j \leq t < t_{j+1}),$$

$$Y_i^j(t) = \mathbb{1}(X_i \geq t, t_j \leq t < t_{j+1})$$

and we consider $t_0 = 0$ and $t_{M+1} = \tau$. If we look more closely, we can see how the counting process $N_i^j(t)$, specific to the intervals $[t_j, t_{j+1})$, is constant outside $[t_j, t_{j+1})$. Hence, its increment, $dN_i^j(t)$, is null for every subjects i that experiences an event outside that specific interval of time. Moreover,

$$Y_i^j(t) = \mathbb{1}(X_i \geq t, t_j \leq X_i < t_{j+1}, t_j \leq t < t_{j+1}) + \mathbb{1}(X_i \geq t_{j+1}, t_j \leq t < t_{j+1}).$$

Therefore, $Y_i^j(t)$ is a function consistently equal to one that becomes null when the subject experiences the event or is censored. Hence, any observation with $X_i < t_j$ doesn't play any role in the j-integral, since $Y_i^j(t) = 0$ and $dN_i^j(t) = 0$. However, if $X_i \geq t_{j+1}$, since $dN_i^j(t) = 0$ and $Y_i^j(t) = 1$, every such observation still appears in the risk set. Indeed, observations that experience the event or are censored after t_{j+1} are still alive in the interval $[t_j, t_{j+1})$ and, therefore, still at risk.

In conclusion, while considering $[t_j, t_{j+1})$interval, we can censor at t_{j+1} anyone that dies after t_{j+1} and we can eliminate anyone that dies or is censored before t_j. More technically, we create therefore for each interval, $[t_0, t_1), \ldots, [t_{M-1}, t_M), [t_M, \infty)$, M+1 working "datasets",

$$D^j = (X_i^j, \delta_i^j, \tilde{Z}_i^j)_{i=1}^{n_j},$$

for $j = 0, \ldots, M$, according to the following principles:

$$X_i^j = \begin{cases} X_i & t_j \leq X_i < t_{j+1} \\ t_{j+1} & X_i \geq t_{j+1} \end{cases},$$

$$\delta_i^j = \begin{cases} \delta_i & t_j \leq X_i < t_{j+1} \\ 0 & X_i \geq t_{j+1} \end{cases}. \tag{2.8}$$

$$\begin{aligned} \tilde{Z}_i^0 &= Z_i(t_0) \\ \tilde{Z}_i^j &= \Big(Z_i(t_j)^\top, \hat{h}_0(\tilde{Z}_i^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_i^{j-1})\Big)^\top. \end{aligned} \tag{2.9}$$

Here, $n_j = |D_j|$, denotes the cardinality of the at-risk observations, i.e., the set D_j. Note that the at-risk datasets, are rarely of the same size and that typically, $n_0 \geq n_1 \geq \cdots \geq n_M$.

The idea described above is inspired by the time-dependent coefficient survival models, utilized widely since the early work on histogram sieves (Murphy and Sen, 1991) or more generally time-varying coefficient models (Hastie and Tibshirani, 1993). The justification can be understood from breaking down the integrated score into a product of time intervals specific score.

2.2.4 Estimation

Each dataset D_j is now used to estimate the part of the risk $h(Z(t),t)$ that is specific to the interval j, that is $h_j(Z(t))$. To this goal, $M+1$ neural networks, one for each time interval, are constructed. To accommodate the sequential nature of the time, observations within D_j together with outcomes of the trained neural networks from previous time intervals, $\hat{h}_k(\tilde{Z}^k)$ for $k<j$, are fed into the neural network j. The neural network j uses, as loss function, the j-th loss $\mathcal{L}_j(\theta_j)$, (2.7), relative to that specific interval. Due to the assumed structure (2.1), that loss simplifies to:

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^{n_j}\int_{t_j}^{t_{j+1}} Y_i^j(t)\left[h_{j,\theta_j}(\tilde{Z}_i^j)-\bar{h}_{j,\theta_j}(t)\right]^2 dt \\ -\frac{1}{n}\sum_{i=1}^{n_j}\int_{t_j}^{t_{j+1}}\left[h_{j,\theta_j}(\tilde{Z}_i^j)-\bar{h}_{j,\theta_j}(t)\right] dN_i^j(t)+\lambda\left\|\theta_j\right\|_p,\end{aligned} \tag{2.10}$$

where

$$\bar{h}_{j,\theta_j}(t)=\frac{\sum_{i=1}^{n_j} h_{j,\theta_j}(\tilde{Z}_i^j)Y_i^j(t)}{\sum_{i=1}^{n_j} Y_i^j(t)}. \tag{2.11}$$

The above function $\bar{h}_{j,\theta_j}(t)$ represents the mean of h_{j,θ_j} restricted to the risk set at time t which comprises all the subjects still alive. Now, noticing that the function $\bar{h}_{j,\theta_j}(t)$ is a stepwise function that is constant on any interval $[X_{r-1}^j, X_r^j]$, as shown in the appendix, the above simplifies to:

$$\frac{1}{2n}\sum_{i=1}^{n_j}\sum_{r=1}^{i}\left[h_{j,\theta_j}(\tilde{Z}_i^j) - \bar{h}_{j,\theta_j}(X_r^j)\right]^2\left(X_r^j - X_{r-1}^j\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n_j}\left[h_{j,\theta_j}(\tilde{Z}_i^j) - \bar{h}_{j,\theta_j}(X_i^j)\right]\delta_i^j + \lambda\left\|\theta_j\right\|_p,$$

where $X_0^j = t_j$. See Appendix 2.6.2 for more details.

Here, we notice how the loss cannot be written as a sum of independent individual i-specific losses. Indeed, the term $\bar{h}_{j,\theta_j}(X_i^j)$, as explained before, uses all the individuals still at risk at time X_i^j. Thus, the optimization method that relies on breaking down the sample in batches cannot be performed here. This is a common characteristic of every loss related to any continuous survival model. It is the same, for example, in Katzman et al. (2018), where the loss used is the partial likelihood that characterizes the Cox proportional model. The application of batch optimization for survival data requires the use in the loss of an approximate risk set, instead of the true one, as explained in Kvamme et al. (2019) where the idea is applied to the Cox model.

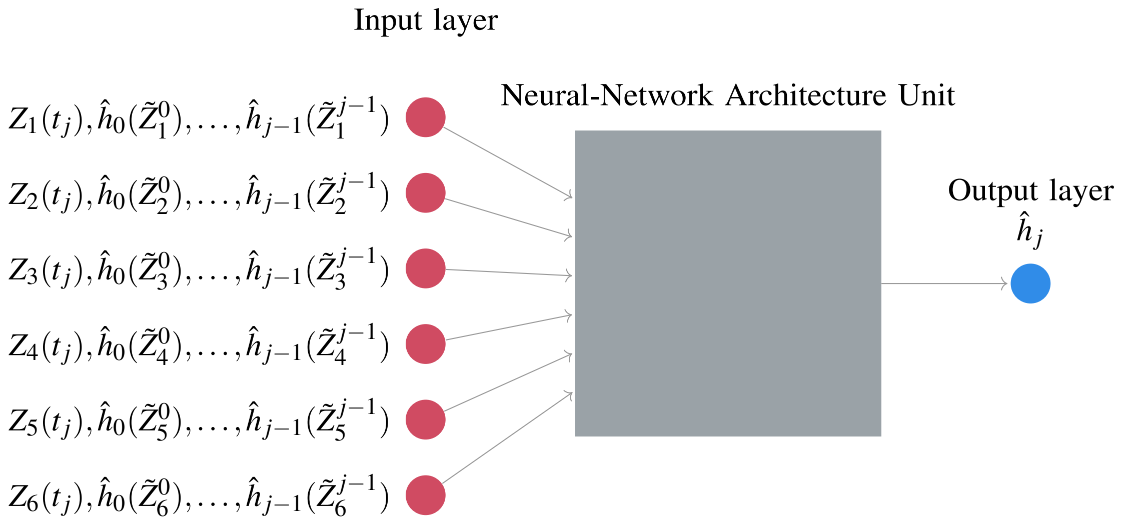

Input layer
Neural-Network Architecture Unit
Output layer
$\hat{h}_j$
$Z_1(t_j), \hat{h}_0(\tilde{Z}_1^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_1^{j-1})$
$Z_2(t_j), \hat{h}_0(\tilde{Z}_2^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_2^{j-1})$
$Z_3(t_j), \hat{h}_0(\tilde{Z}_3^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_3^{j-1})$
$Z_4(t_j), \hat{h}_0(\tilde{Z}_4^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_4^{j-1})$
$Z_5(t_j), \hat{h}_0(\tilde{Z}_5^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_5^{j-1})$
$Z_6(t_j), \hat{h}_0(\tilde{Z}_6^0), \ldots, \hat{h}_{j-1}(\tilde{Z}_6^{j-1})$

Figure 2.2: Deep Hazard: An example of j-th NN with Input Layer consisting of six still at risk observations and forward passes of previously learned $j-1$ networks, with arbitrary NN Architecture Unit.

In our experiments, every hidden fully-connected layer is followed by a nonlinear activation function and a dropout layer; however, the method can take any architecture of the layers of interest. As is common in neural networks, the output is a plain weighted combination of the last hidden layer's output. No activation function is used for the computation of the output. An example of the input structure of each j-th NN network is depicted in Figure 2.2. We observe how each set of still-at-risk observations is enriched with additional features coming out of feed-forward passes run on previously fitted $j-1$ networks.

A possible time-convolution, as proposed earlier, can be visualized in Figure 2.3. There, we illustrate how the time-dependent outputs of each j-th NN are passed onto all of the future NNs. Moreover, we indicate that each dataset, D_j, comprising inputs of j-th NN, depends on the previous dataset D_{j-1}.

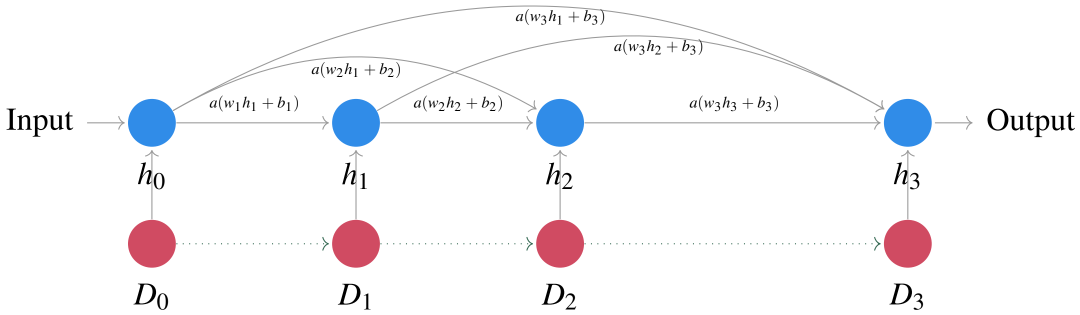


Figure 2.3: Deep Hazard: unpacked time-convolutions. Each time-specific neural network (NN) disregards hidden layers. The blue nodes denote outputs of time-specific NNs whereas, arrows denote feed-forward interactions over time. At each arrow, we show the activation function a and weights w_i and biases b_i, $i = 1,2,3.$. Dotted lines denote dependence of still-at-risk individuals comprising inputs of each time-specific NN.

After running $M+1$ sequential (or time-convoluted) neural networks on the $M+1$ working datasets, we obtain the optimized weights and biases, denoted here with $\theta_0,\ldots,\theta_M$. To form prediction of the new, test individual we proceed as follows. With a little abuse in notation, let

$$Z(t) = \{Z(t_0),\ldots,Z(t_M)\},$$

be observations pertaining to a sequence of follow up visits, at times $t_0,\ldots,t_M$, of a new patient. For each $\theta_0,\cdots,\theta_M$, with standard forward pass (evaluating the estimated hazard for a specific new observation), one for each network, we get the following estimates:

$$\hat{h}_0(\tilde{Z}^0) := h_{\theta_0}(\tilde{Z}^0),\ldots,\hat{h}_M(\tilde{Z}^M) := h_{\theta_M}(\tilde{Z}^M).$$

Here, each $\tilde{Z}$ is constructed following equation (2.9), in other words, each prediction of the risk at a future time t_j uses the formed predictions of previous time points $t_0,\cdots,t_{j-1}$. Combining these into a single risk estimator is then simple. Following (3.1) we obtain for an out-of-sample individual

$$\hat{h}(Z(t),t) = \hat{h}_0(\tilde{Z}^0)\mathbb{1}(t \le t_1) + \hat{h}_1(\tilde{Z}^1)\mathbb{1}(t_1 < t \le t_2) + \cdots + \hat{h}_M(\tilde{Z}^M)\mathbb{1}(t > t_M). \tag{2.12}$$

For more details on the training process see Algorithm 1.

Algorithm 1 DeepHazard:Training

```
Require: Training set $(X_i, \delta_i, Z_i(t_0), \dots, Z_i(t_M))_{i=1}^n$, hyper parameters and hidden layers of $M+1$
  neural networks
  $\theta \leftarrow$ initialize weights and biases
  Set $\tilde{Z}_i^0 \leftarrow Z_i(t_0)$
  Create $D_0$ dataset according to (2.8)
  $\theta_0 \leftarrow$ neural network initialized at $\theta$ and with input $\tilde{Z}_1^0, \dots, \tilde{Z}_n^0$
  for $j$ in $0 : M$ do
      $\theta \leftarrow$ initialization of weights and biases accordingly to initialization method
      Set $\tilde{Z}_i^j \leftarrow \Big(Z_i(t_j), \hat{h}_0(\tilde{Z}_i^0), \dots, \hat{h}_{j-1}(\tilde{Z}_i^{j-1})\Big)$
      Create $D_j$ dataset according to (2.8)
      Set $n_j = \mathrm{card}(D_j)$
      $\theta_j \leftarrow$ neural network initialized at $\theta$ and with input $\tilde{Z}_1^j, \dots, \tilde{Z}_{n_j}^j$
      for $i$ in $1 : n$ do
          $\hat{h}_j(\tilde{Z}_i^j) \leftarrow h_{\theta_j}(\tilde{Z}_i^j)$        ▷ forward pass of the $j$-th NN on the training data
          $r_i \leftarrow \sum_{l=1}^n Y_l(X_i)$        ▷ number of people at risk at that time
          $\bar{h}_j(X_i) \leftarrow \sum_{l=i}^n r_i^{-1} \hat{h}_j(\tilde{Z}_l^j)$
  for $i$ in $1 : n$ do
      $J_i \leftarrow \{ j \, : \, t_j \le X_i < t_{j+1} \}$        ▷ which interval contains the censored time
      $r_i \leftarrow \sum_{l=1}^n Y_l(X_i)$

          $\hat{\Lambda}_0(X_i) \leftarrow \sum_{l=1}^{i} \frac{\delta_l}{r_l} - \sum_{j=0}^{J_i} \sum_{s:t_j \le X_s < t_{j+1}} [X_{s+1} - X_s]\bar{h}_j(X_s)$

      return
      A matrix $\hat{h} = [\hat{h}_j(\tilde{Z}_i^j)]_{i=1,\dots,n;j=1,\dots,M}$.
      The vectors $\theta_0, \dots, \theta_M$        ▷ weights and biases for each neural network
      A vector $\Big(\hat{\Lambda}_0(X_1), \dots, \hat{\Lambda}_0(X_n)\Big)$
```

2.2.5 Prediction of the survival for external covariates

Practitioners are often concerned with predicting the survival rate of a new patient for a given period of time in the future: survival at one, five, twenty years after diagnosis, for example. Time-dependent covariates may be classified as external and internal. The former are covariates that are fixed or whose total path is determined in advance for each individual under study, while the latter are covariates whose values are generated by the individual. While survival prediction may be performed for the former, since internal covariates carry information about the failure time, it is never carried out for the latter (Kalbfleisch and Prentice, 2011). In the following we explain how to use our method to predict the survival function for time-dependent external covariates.

With estimated risks of the previous section, we only need to design baseline estimates of the hazard. When considering an additive hazards model $\lambda(t|D,Z) = \lambda_0(t) + \beta Z(t)$ as explained in Lin and Ying (1994a), a semiparametric estimate of the cumulative baseline can be proposed. Here, we directly extend their semi-parametric approach.

Observe that under the model (3.1),

$$dN_i(t) = dM_i(t) + \int_0^t Y_i(u) d\Lambda_0(u) + \int_0^t Y_i(u) h(Z_i(u), u) du,$$

it is natural to consider the following estimator

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n \left\{ dN_i(u) - Y_i(u)\hat{h}(Z_i(u), u) du \right\}}{\sum_{i=1}^n Y_i(u)}, \tag{2.13}$$

with $\hat{h}$ as defined in (2.12). Our time-convolutions provide a way to also estimate cumulative baseline hazards for each time-interval. Therefore our method allows data exploration in each time interval as well as overall. We show the equivalence of the two approaches in the Appendix 2.6.3.

Lastly, we compute the predicted survival curve by combining the results of the M neural network predictions in the following way

$$\hat{S}(t \mid Z(t)) = \begin{cases} \exp(-\hat{\Lambda}_0(t) - \hat{h}_0(\tilde{Z}^0)t) & t < t_1 \\ \exp(-\hat{\Lambda}_0(t) - \hat{h}_1(\tilde{Z}^1)(t-t_1) - \hat{h}_0(\tilde{Z}^0)t_1) & t_1 \leq t < t_2 \\ \exp(-\hat{\Lambda}_0(t) - \hat{h}_2(\tilde{Z}^2)(t-t_2) - \hat{h}_1(\tilde{Z}^1)(t_2-t_1) - \hat{h}_0(\tilde{Z}^0)t_1) & t_1 \leq t < t_2 \\ \dots \end{cases} . \quad (2.14)$$

Here, we take as value for $Z(t) = Z(t_{J_t})$ where $J_t = \{j \,:\, t_j \leq t < t_{j+1}\}$.

Finally we construct the following adjusted version of the predicted survival

$$\hat{S}(t \mid Z(t)) = \min_{s \leq t} \hat{S}(s \mid Z(s)),$$

guaranteeing the estimator of the survival to be decreasing, consequently avoiding the well known problem of possibly negative risk and therefore hazard. For more details see Algorithm 2.

Algorithm 2 DeepHazard:Prediction

```
Require: Test data $\mathbb{Z}(t_0),\dots,\mathbb{Z}(t_M)$, event times from the Test data, $\{X_1,\dots,X_n\}$ and outcomes of
    DeepHazard:Training, i.e., $\theta_0,\dots,\theta_M$ and $\hat{\Lambda}_0(X_1),\dots,\hat{\Lambda}_0(X_n)$
    Set $\tilde{\mathbb{Z}}^0 \leftarrow \mathbb{Z}(t_0)$
    for $j$ in $0:M-1$ do
        $\hat{h}_j(\tilde{\mathbb{Z}}^j) \leftarrow h_{\theta_j}(\tilde{\mathbb{Z}}^j)$        ▷ forward pass of the DeepHazard on the test data
        $\tilde{\mathbb{Z}}^{j+1} \leftarrow \big(\mathbb{Z}(t_{j+1}), \hat{h}_0(\tilde{\mathbb{Z}}^0),\dots,\hat{h}_j(\tilde{\mathbb{Z}}^j)\big)$
    $\hat{h}_M(\tilde{\mathbb{Z}}^M) \leftarrow h_{\theta^M}(\tilde{\mathbb{Z}}^M)$
    for $i$ in $1:n$ do
        Set $J_i \leftarrow \{j \,:\, t_j \leq X_i < t_{j+1}\}$        ▷ Here $\mathbb{Z}(X_i) \approx \mathbb{Z}(t_{J_i})$
        Set
            $$\hat{S}(X_i \mid \mathbb{Z}(X_i)) \leftarrow \exp\left(-\hat{\Lambda}_0(X_i) - \hat{h}_{J_i}(\tilde{\mathbb{Z}}^{J_i})(X_i - t_{J_i}) - \sum_{l=0}^{J_i-1} \hat{h}_l(\tilde{\mathbb{Z}}^l)(t_{l+1}-t_l)\right)$$
        Set $\hat{S}(X_i \mid \mathbb{Z}(X_i)) \leftarrow \min_{l \leq i} \hat{S}(X_l \mid \mathbb{Z}(X_l))$        ▷ monotonicity guarantee
     return $\hat{S}(X_1 \mid \mathbb{Z}(X_1)),\dots,\hat{S}(X_n \mid \mathbb{Z}(X_n))$
```

Until now, we have implicitly assumed that, any new observation will have $Z(t)$ measured at the same time points used to train the network - $t_0,\dots,t_M$. If this is not the case we approximate $Z(t_j)$ with the nearest $Z(t)$ available. More in details, if the measurements $Z(\tilde{t}_0),\dots,Z(\tilde{t}_{\tilde{M}})$ are available, we make use of the following approximation: $Z(t_i) = Z(\tilde{t}_{J_i})$ where $J_i = \arg\min_{j=1,\dots,\tilde{M}} |t_i - \tilde{t}_j|$.

We study in simulation the effect of the number and the placement of the time points that define the $M+1$ intervals. We show that the performance of our procedure remains stable when the time points at which the covariates are measured shift or more time points are added. The only restriction that needs to be kept in mind is that we need to have enough observations to train the last neural network, that, we remind, uses as input only the observation still at risk after t_M. The last time point therefore cannot be too large in comparison to the magnitude of the censored event time of our sample.

2.3 Finite sample experiments

In this section, we evaluate the performance of Deep Hazard in finite samples. We compare DeepHazard with the Additive Hazards Model, (Aalen, 1980), that presuposes

$$\lambda(t \mid Z(t)) = \lambda_0(t) + \beta(t)Z(t),$$

and with the Time dependent Cox Model, (Fisher and Lin, 1999), that assumes

$$\lambda(t \mid Z(t)) = \lambda_0(t) \exp\left(\beta Z(t)\right).$$

We use the R packages *Timereg* and *Survival*, respectively to fit the Additive Hazards Model and the Time dependent Cox model. As a measure of performance, we use the time dependent C-index as proposed by Antolini et al. (2005),

$$\mathrm{C}_{\text{index}} = \sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} \mathbb{1}\left(\hat{S}(t_i \mid Z_i(t)) < \hat{S}(t_j \mid Z_j(t))\right) p_{i,j},$$

where

$$p_{i,j} = \frac{\mathbb{1}\left\{t_i < t_j,\ \delta_i = 1\right\} + \mathbb{1}\left\{t_i = t_j,\ \delta_i = 1,\ \ \delta_j = 0\right\}}{\sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} \left[\mathbb{1}\left\{t_i < t_j,\ \delta_i = 1\right\} + \mathbb{1}\left\{t_i = t_j,\ \delta_i = 1,\ \ \delta_j = 0\right\}\right]}.$$

We also introduce a new measure, the integrated mean square prediction error (IMSPE), defined as follows:

$$\mathrm{IMSPE} = \frac{1}{\tau} \int_0^\tau \frac{1}{n} \sum_{i=1}^{n} \left\{\hat{S}(t \mid Z(t)) - S(t \mid Z(t))\right\}^2,$$

to capture the quality of the prediction error through time.

We implement our neural network in PyTorch, `https://github.com/deniserava/DeepHazard`. The implementation is flexible in that the user can choose the structure of the Neural Network: the number of hidden layers, number of hidden nodes, activation function, and a dropout rate. Moreover, the following Hyperparameters related to the optimization procedure of the neural networks, as initialization method, optimizer used, learning Rate (lr), number of Epochs (E), and early stopping can be chosen. The user can also select the regularization parameters λ and p of the loss (2.10). It is worth noting that Epochs are updating the network weights and biases (parameters) through a suitable optimization method, but stay un-permuted to preserve the order of the survival outcomes. A list of the popular activation functions, that we implemented, can be found in Appendix 2.6.1; see Table 2.15.

Our numerical experiments are evaluated on simulated data. We focus on the settings with time-varying covariates. There is a need to describe data-generating processes for the hazard models in the presence of time-varying covariates. The latter are generated using the procedure described in Algorithm 3.

2.3.1 Impact of the sample size

We assume the data is generated according to the following four different hazards models. Below '$*$' denotes multiplication. Model 1 follows additive structure but the covariates are highly correlated and non-linear. Model 2 considers further interactions with time whereas Model 3 works with highly non-linear interactions. Model 4 is perhaps the most challenging one.

Model 1:

$$\lambda(t \mid Z) = 4t^3 + Z_1(t) * Z_2(t) + Z_1(t) * Z_3(t) + Z_1(t) * Z_3(t) * Z_2(t).$$

Algorithm 3 Time-dependent Simulation

```
Require: Covariate function $z$ such that $Z(t) = z(Z,t)$ and $Z$ follows distribution $\mathcal{Z}$, hazard function
  $h(\cdot,\cdot)$, baseline hazard $\lambda_0(\cdot)$, censoring level $\ell$, follow up times $t_1,\ldots,t_M$, sample size $n$
  for $i = 1,\ldots,n$ do
      Simulate $\omega$ from Uniform distribution $U(0,1)$
      Let $Z$ be a realization of a random draw from $\mathcal{Z}$.
      Let $T_i = t$ where $t$ solves
                                  $f(t) = \omega$
        where $Z(u) := z(Z,u)$
                  $f(t) := \exp\left[-\int_0^t \{\lambda_0(u) + h(u,Z(u))\}\,du\right]$
                                            ▷ $f(t)$ stands to denote the function $S(t|Z(t))$
      for $j = 1,\cdots,M$ do
          Let $Z_i(t_j) = z(Z,t_j)$
  Simulate $n$ independent censoring time $C_i$ from Uniform distribution $U(0,c)$
                                  ▷ $c$ is such that censoring level is below some level $\ell$
  Set $X = \min\{T,C\}$
                                                         ▷ Observed censored event times
  Let $\delta = \mathbb{1}\{T \le C\}$                      ▷ Observed censoring indicator
      return Data $\{X_i,\delta_i,Z_i(t) := (Z_i(t_1),\ldots,Z_i(t_M))\}_{i=1}^n$
```

Model 2:

$$\lambda(t \mid Z) = 4t^3 + \cos(t)[Z_1(t) * Z_2(t)] + |\log(t+1)|Z_1(t) * Z_2(t) + t^3 Z_3(t)^2.$$

Model 3:

$$\begin{aligned}\lambda(t \mid Z) = 4t^3 + \cos(t)[Z_1(t) * Z_2(t)] + \mid \log(t+1) \mid Z_1(t)Z_2(t) + t^3 Z_3(t)^2 \\ + \cos[Z_1(t) * Z_3(t)] + Z_1(t) * Z_3(t) + \frac{1+t^2}{t+1} Z_1(t) * Z_2(t) + Z_1(t)^3 * Z_2(t)^4\end{aligned}$$

Model 4:

$$\lambda(t \mid Z) = 4t^3 + \frac{1}{t+1} Z_1(t) * Z_2(t) + \frac{1}{Z_1(t) * Z_2(t) * Z_3(t)^2 + 1}.$$

The covariates are generated according to the following structure

$$Z_i(t) = \begin{cases} \sqrt{t}\, Z_{0i} & t \leq 0.6 \\ \sqrt{0.6}\, Z_{0i} & \text{otherwise} \end{cases} \tag{2.15}$$

where $i = 1,2,3$ and $Z_{0i} \sim U(0,20)$ for $i = 1,2,3$ except for **Model 1**, where $Z_{01} \sim U(0,10), Z_{02} \sim U(0,20), Z_{0i} \sim U(0,30)$.

We assume to measure the covariates at the following times $0.001, 0.2, 0.4, 0.6$. We generate 1000 observations for the training set and for the test set. We fit to the training set the Additive Hazards Model, the Time-dependent Cox Model and DeepHazard. 1000 epochs are used with early stopping rate $1e^{-5}$ and initialization method he Normal is employed. The C-index of each Model is presented in Table 2.1. The Hyperparameters chosen for our neural network are reported in Table 2.2.

We report also the Oracle C-index that uses the true $S(t \mid Z(t))$ for comparison purposes. We then repeat the simulations with a sample size of 200 for both train and test set. We observe superior performance of DeepHazard both across samples as well as Models. Moreover, C-index is often extremely close to the oracle C-index indicating certain optimality.

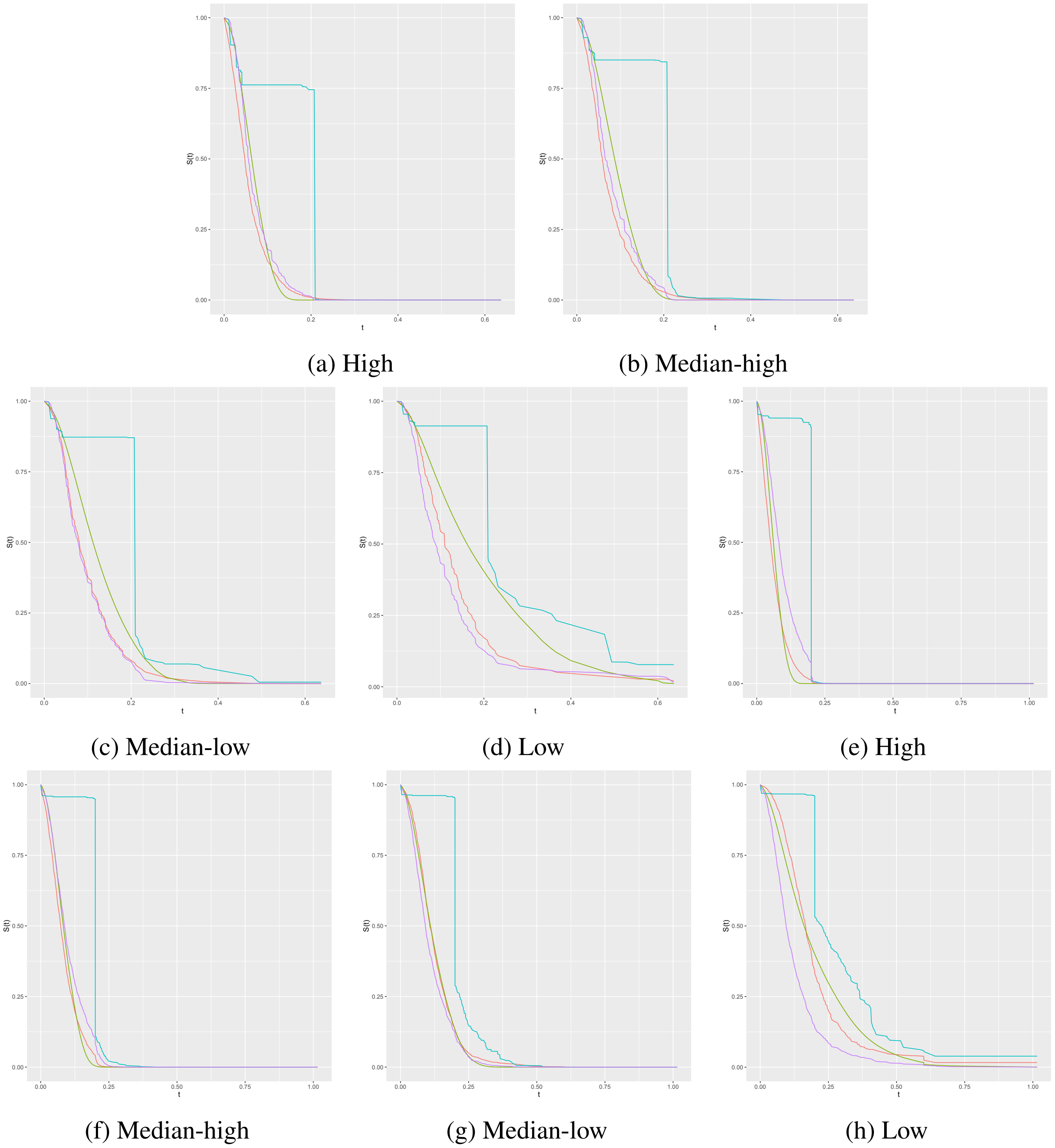
Table 2.1: Result of Simulation for additive Hazards Model, Time-dependent Cox and our method (DeepHazard) for Model 1, 2 , 3 and 4.

	Model 1	Model 2	Model 3	Model 4
C-index				
$n = 1000$				
Oracle	0.765	0.749	0.716	0.742
Deep Hazard	0.752	0.735	0.716	0.733
Additive Hazards	0.665	0.590	0.674	0.636
Time-dependent Cox	0.726	0.718	0.703	0.717
$n = 200$				
Oracle	0.743	0.734	0.681	0.739
Deep Hazard	0.726	0.717	0.666	0.727
Additive Hazards	0.635	0.174	0.651	0.598
Time-dependent Cox	0.713	0.700	0.676	0.699

Table 2.2: DeepHazard experimental Hyperparameters of Table 2.1.

Hyperparameter				
$n = 1000$	Model 1	Model 2	Model 3	Model 4
Optimizer	*Adam*	*Adam*	*Adam*	*Adam*
Activaction	*Elu*(0.1)	*Relu*	*Elu*(0.1)/*Selu*	*Selu*
N. Dense Layer	5	2	2	2
N. Nodes Layer	10/15/20/15/10	10	20	10
Learning rate	0.01	$2e-2$	$2e-1$	$2e-1$
λ	$1e-5$	$1e-3$	$1e-5$	$1e-5$
Penalty	Ridge	Ridge	Ridge	Ridge
Dropout	0.2	0.2	0.2	0.2
$n = 200$				
Optimizer	*Adam*	*Adam*	*Adam*	*Adam*
Activaction	*Selu*	*Relu*	*Selu*	*Relu*
N. Dense Layer	2	2	3	2
N. Nodes Layer	10	10	10/15/10	10
Learning rate	$2e-1$	$2e-2$	$1e-3$	$2e-1$
λ	$1e-2$	0.41	0.61	$1e-4$
Penalty	Ridge	Ridge	Ridge	Ridge
Dropout	0.2	0.2	0.1	0.2

For **Model 3**, both for small and large sample, we plot in Figure 2.4, the true and the estimated survival functions by the three different methods. We divide observations into high, median-high, median-low and low risk according to the risk value $\sum_{j=1}^{4} h_j(Z(t))/4$, i.e., the mean of the all interval specific risk scores $h_j(Z(t))$. We observe a strong bias of the Additive Hazards Model despite a low C-index value. It is often very far from the true survival function. Figure 2.4 part (c) illustrates that $\hat{S}_{AddHaz}(0.19 \mid Z(0.19)) \approx 0.875$ while the true survival function satisfies $S(0.19 \mid Z(0.19)) \approx 0.187$. On the other hand $\hat{S}_{DeepHaz}(t \mid Z(t))$ is a good smooth approximation of the true function. We also observe that larger samples lead to a better Deep Hazard approximation.



(a) High

(b) Median-high

(c) Median-low

(d) Low

(e) High

(f) Median-high

(g) Median-low

(h) Low

Figure 2.4: Survival curves for Model 3 for different groups of subject for $n = 200$ (a)-(d) and $n = 1000$ (e)-(h). Red color denotes the proposed DeepHazard, Blue denotes the time-varying Additive Hazards Method, Green denotes the true Survival curve and Purple denotes the Time-dependent Cox.

Further studies on **Model 3** were done to showcase the impact of the architecture on the learning. We see that our procedure preforms better both in terms of C-index as well as IMSPE measure of prediction quality. We see that DeepHazards is showcasing IMSPE improvement from 50% to 200%.

Table 2.3: Model 3 where each Layer is dense and learning rate is $2e-1$ unless specified differently. Activation function is Relu and $\lambda = 1e-5$ with Ridge penalty. 'lr' stands for learning rate, 'Deep Haz' stands for Deep Hazard, 'Add Haz' stands for Additive Hazard, 'TV Cox' stands for Time-varying Cox

	Architecture										
# of Layers	One		Two				Three		Four		Ten
Node x layer	[50]	[50]	[10]	[10]	[50]	[50]	[10]	[10]	[10]	[10]	[10]
lr		$2e-2$		$2e-2$		$2e-2$		$2e-2$		$2e-2$	$2e-2$
IMSPE $*100$											
Deep Haz	0.311	0.282	0.365	0.287	0.369	0.423	0.409	0.316	0.315	0.326	0.529
Add Haz	7.373	7.373	7.373	7.373	7.373	7.373	7.373	7.373	7.373	7.373	7.373
TV Cox	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967	0.967
C-index											
Deep Haz	0.717	0.708	0.710	0.714	0.723	0.744	0.705	0.796	0.696	0.710	0.695
Add Haz	0.674	0.674	0.674	0.674	0.674	0.674	0.674	0.674	0.674	0.674	0.674
TV Cox	0.683	0.683	0.683	0.683	0.683	0.683	0.683	0.683	0.683	0.683	0.683
Oracle	0.716	0.716	0.716	0.716	0.716	0.716	0.716	0.716	0.716	0.716	0.716

Lastly, we investigated the impact of the activation functions. Setting is that of **Model 3** with Two Layers each comprised of ten (dense) nodes. Learning rate was fixed at $2e-1$.

Table 2.4: Results within Model 3 across different activation functions

	Relu	*Selu*	*Atan*	*Tanh*	*LogLog*	*LeakyRelu*
IMSPE $*100$						
Deep Hazard	0.269	0.238	0.308	0.353	0.298	0.399
Additive Hazards	7.373	7.373	7.373	7.373	7.373	7.373
Time-varying Cox	0.967	0.967	0.967	0.967	0.967	0.967
C-index						
Deep Hazard	0.709	0.705	0.709	0.688	0.692	0.705
Additive Hazards	0.674	0.674	0.674	0.674	0.674	0.674
Time-varying Cox	0.683	0.683	0.683	0.683	0.683	0.683
Oracle	0.716	0.716	0.716	0.716	0.716	0.716

2.3.2 Impact of a large number of time-varying covariates

We assume the data is generated according to the following different models:

- Model 5:

$$\begin{aligned}\lambda(t \mid Z) = 4t^3 + \cos(t)[Z_1(t) * Z_2(t)] + |\log(t+1)|Z_1(t) * Z_2(t) \\ + t^3 Z_3(t)^2 + \frac{1}{1 + Z_{20}(t) * Z_1(t) + \sqrt{t}}.\end{aligned}$$

- Model 6:

$$\begin{aligned}\lambda(t \mid Z) = 4t^3 + \cos(t)[Z_1(t) * Z_2(t)] + |\log(t+1)|Z_3(t) * Z_4(t) + t^3 Z_5(t)^2 \\ + \cos[Z_6(t) * Z_7(t)] + Z_8(t) * Z_9(t) + \frac{1+t^2}{t+1} Z_{10}(t) * Z_{11}(t) \\ + Z_{12}(t)^3 * Z_{13}(t)^4 + \frac{1}{1 + Z_{20}(t) * Z_{14}(t) + \sqrt{t}}.\end{aligned}$$

where $Z_i(t)$ follows (2.15).

For **Model 5** all Z_{0i} are drawn from $U(0,20)$ except for Z_{01} that is drawn from $U(5,20)$

and Z_{020},Z_{019} from $U(3,4)$ and Z_{16},Z_{17},Z_{18} from $U(0,1)$. For **Model 6** all Z_{0i} drawn from $U(0,20)$ except for Z_{01} from $U(5,20)$ and Z_{020},Z_{019},Z_{04} from $U(3,4)$ and Z_{16},Z_{17},Z_{18} from $U(0,1)$.

We considered the following measurement times $0.001, 0.2, 0.4, 0.6$ for **Model 5** and at $0.001, 0.1, 0.2, 0.3$ for **Model 6**. We generate 1000 observation for the training set and for the test set. The Hyperparameters chosen for our neural network are reported in Table 2.6. 1000 epochs are used with early stopping rate $1e^{-5}$ and initialization method he Normal is employed. The C-index of each model is presented in Table 2.5. In these cases we observe strong failure of the additive hazards model with C-index being extremely low, especially for non-linear time interactions. Time-varying Cox approach had difficulties due to the periodic covariate effects.

Table 2.5: Results of Simulation for Additive Hazards Model, Time-dependent Cox and our method (DeepHazard) for Model 5 and 6.

C-index		
$n = 1000$	Model 5	Model 6
Deep Hazard	0.691	0.635
Additive Hazards	0.135	0.423
Time-varying Cox	0.677	0.598

Table 2.6: DeepHazard experimental Hyperparameters for Model 5 and 6.

Hyperparameter	Model 5	Model 6
Optimizer	*Sgd*	*Adam*
Activaction	*Elu*(0.1)	*Selu*
N. Dense Layer	1	1
N. Nodes Layer	20	20
Learning rate	$2e-1$	$2e-1$
λ	0.56	0.1
Penalty	Ridge	Ridge
Dropout	0.2	0.2

2.3.3 Impact of the censoring rate

We assume the data is again generated according to the **Model 4**, with $Z_i(t)$ following (2.15). and $Z_{0i} \sim U(0,20)$ for $i = 1,2,3$. We assume to measure the covariates at the following times $0.001, 0.2, 0.4, 0.6$. We generate our data under different censoring scenario: $10\%, 20\%$. We also consider the setting of **Model 5** and **Model 6** with covariates measured at $0.001, 0.1, 0.2, 0.3$ and $0.001, 0.1, 0.15, 0.2$, respectively, each with censoring of $0\%, 15\%$, and 30%.

We generate 1000 observations for the training set and for the test set. The Hyperparameters chosen for our neural network are reported in Table 2.8. 1000 epochs are used with early stopping rate $1e^{-5}$ and he-Normal initialization. The C-index of each model is presented in Table 2.7. The result shows strong stability with respect to censoring.

Table 2.7: C-index for additive Hazards Model, Time-dependent Cox and our method (DeepHazard) under different censoring scenarios.

C-index								
	Model 4		Model 5			Model 6		
Censoring	10%	20%	0%	15%	30%	0%	15%	30%
DeepHazard	0.724	0.719	0.682	0.678	0.681	0.641	0.632	0.623
Additive Hazards	0.532	0.625	0.504	0.501	0.413	0.506	0.498	0.417
Time-dependent Cox	0.713	0.699	0.676	0.671	0.674	0.604	0.592	0.598

Table 2.8: DeepHazard experimental Hyperparameters

Hyperparameter	Model 4	Model 5	Model 6
Censoring	$(10\%, 20\%)$	$(0\%, 15\%, 30\%)$	(0%,15%,30%)
Optimizer	*Adam*	*Sgd*	*Sgd*
Activation	*Selu*	*Elu*(0.7)	*Elu*(0.5)
N. Dense Layer	$2, 1$	2,2,3	2
N. Nodes Layer	10, 20	20	20
Learning rate	$2e-1, 3e-3$	$1e-2, 1e-2, 1e-1$	$1e-2$
λ	$1e-5, 1e-4$	$0.061, 0.061, 0.05$	0.061
Penalty	Ridge	Lasso	Lasso
Dropout	0.2	$0.1/0.15, 0.1/0.15, 0.1/0.15/0.15$	$0.1/0.15$

Further studies on the impact of the censoring and architecture structure were performed under **Model 2**. We worked with 1000 samples in the training and testing phase and report the findings in Table 2.9.

Table 2.9: C-index and IMSPE for Deep Hazard, Additive Hazards and Time-dependent Cox model for Model 2. For the Deep Hazard, the dropout rate is 0.2 and $\lambda = 1e-3$ with Ridge penalty and Adam optimizer is used. Architecture, activation function and learning rate is specified in the table.

Architecture	10/10, Relu, $2e-2$		20/20, Leaky Relu, $2e-3$	
Censoring	0%	10%	0%	10%
		IMSPE$*$10		
Deep Hazard	0.045	0.051	0.036	0.048
Additive Hazards	0.641	0.785	0.641	0.785
Time-dependent Cox	0.249	0.340	0.249	0.340
		C-index		
Deep Hazard	0.735	0.746	0.732	0.743
Additive Hazards	0.592	0.102	0.592	0.102
Time-dependent Cox	0.718	0.707	0.718	0.707

2.3.4 Effect of shifting the time points at which the covariates are measured

We assume the data is again generated according to the **Model 6**. The covariates are assumed to be measured at the following different sets of time points:

(A) $0.001, 0.1, 0.15, 0.2$; (B) $0.001, 0.05, 0.08, 0.12$; (C) $0.001, 0.15, 0.2, 0.25$;
(D) $0.001, 0.05, 0.08, 0.12, 0.15, 0.2$.

We generate 1000 observation for the training set and for the test set. The Hyperparameters chosen for our neural network are reported in Table 2.11. 1000 epochs are used with early stopping rate $1e^{-5}$ and initialization method he Normal. The C-index of each Model is presented in Table 2.10. Our methods outperforms the other traditional ones for every sets of time points. Moreover, it is interesting to notice how, while the C-index of Ls and Cox depends on where the covariates are measured, our method presents greater stability with respect to the shift. Our C-index is indeed roughly always 0.63 no matter at which and how many time points the measurements are taken.

Table 2.10: Results of Model 6 for additive Hazards Model, Time-dependent Cox and our method (DeepHazard) for different censoring scenario.

C-index				
	(A)	(B)	(C)	(D)
DeepHazard	0.633	0.630	0.633	0.632
Additive Hazards	0.506	0.572	0.485	0.605
Time-dependent Cox	0.604	0.620	0.601	0.619

Table 2.11: DeepHazard experimental Hyperparameters

Time points	A	B	C	D
Hyperparameters				
Optimizer	*Adam*	*Adam*	*Adam*	*Adam*
Activaction	$Elu(0.5)$	$Elu(0.5)$	$Elu(0.5)$	$Elu(1.5)$
N. Dense Layer	2	2	2	2
N. Nodes Layer	20	20	20	20
Learning rate	$1e-2$	$1e-2$	$1e-2$	$1e-2$
λ	0.061	0.0007	0.08	0.0001
Penalty	Lasso	Lasso	Lasso	Lasso
Dropout	$0.1/0.15$	$0.1/0.15$	$0.1/0.15$	$0.1/0.15$

2.4 Real data experiments

In this section we use our method on three benchmark real datasets.

We compare our method with semiparametric additive hazards Model that assumes:

$$\lambda(t \mid Z) = \lambda_0(t) + \beta Z,$$

survival random forest, (Ishwaran et al., 2008), as well as Deepsurv of Katzman et al. (2018). Deepsurv is a Cox proportional hazards deep neural network that assumes proportionality of the hazard but it doesn't assume linearity of the risk as the standard Cox model:

$$\lambda(t \mid Z) = \lambda_0(t)\exp\{h(Z)\}.$$

We use the R package *Timereg* and the Python package *PySurvival*, respectively to fit the Additive Hazards Model and DeepSurv.

Notice that both DeepSurv and the traditional Cox Model rely on the proportional hazard assumption, under which the ratio of the cumulative hazards between groups is assumed to be constant with time. As a diagnostic, for each of the dataset analyzed, we plot this ratio between groups defined by binary covariates. Not constant line in this type of plot indicates departure from the proportional hazard assumption; see Figure 2.5.

With slight abuse in notation, as a measure of predictive capability of the models, we report the traditional concordance index, defined as

$$\mathrm{C}_{\text{index}} = \frac{\sum_{i,i'} \mathbb{1}(X_i > X_{i'})\mathbb{1}(h(Z_i) < h(Z_{i'}))\delta_{i'}}{\sum_{i,i'} \mathbb{1}_{X_i > X_{i'}}\delta_{i'}}.$$

2.4.1 Molecular Taxonomy of Breast Cancer International Consortium dataset (METABRIC)

The dataset consists of gene expression and clinical features for 1980 breast cancer patients,(Curtis et al., 2012). The time variable is time to death and 57.72% of observations experienced the event. For ease of comparison we use, as training and test set, the same dataset used in Katzman et al. (2018) where 20% of the data are saved as test set. As covariates 4 gene indicators are used plus hormone treatment indicator, radiotherapy indicator, chemotherapy indicator, ER-positive indicator and age at diagnosis.

We report in Table 2.12 the C-index for us, DeepSurv, Semiparametric Additive Hazards Model (LS) and Survival Random Forest. For our Neural Netwok we use one layer with 40 nodes, Elu activaction function with $\alpha = 0.1$, Adam optimizer, learning rate of 0.001, $\lambda = 1e-4$ for Ridge penalty and 0.1 for Dropout. For DeepSurv we use the hyperparameters reported in their paper. One layer with 41 nodes, Selu activaction function, Adam optimizer, learning rate of 0.010, $\lambda = 10.891$ for Ridge penalty and 0.160 as Dropout rate.

In Table 2.12, in parenthesis, we write the result reported by Katzman et al. (2018) for both DeepSurv and RSF. We plot in Figure 2.5a the ratio of the cumulative hazards between four groups defined by the four patient's clinical features (hormone treatment indicator, radiotherapy indicator, chemotherapy indicator, ER-positive indicator).

It is clear from the plot how these ratios are not constant with time and therefore how the proportional hazards assumption, on which Deepsurv is based, is violated. From the results, our method indeed outperforms Deepsurv. Moreover it outperforms random survival forest which we fine-tuned. RSF C-index, as per tuning, was very comparable with Deep Surv.

Table 2.12: Results for Metabric dataset Results in parenthesis are the reported numbers of Katzman et al. (2018) of the corresponding methods.

C-index	
Deep Hazard	**0.664**
Additive Hazards	0.645
Deep Surv	0.650 (0.654)
RSF	0.647 (0.619)

2.4.2 Rotterdam and German Breast Cancer Study Group dataset (GBSG)

The dataset consists of 1546 patients with node-positive breast cancer (Schumacher et al., 1994). The time variable is time to death and 90% experienced the event. Again, as training and test set, we use the same dataset used in Katzman et al. (2018) where 20% of the data are saved as test set. The testing data consists of 686 patients in a randomized clinical trial that studies the effect of chemotherapy and hormone treatment on survival rate. We report in Table 2.13 the C-index for us, DeepSurv, Semiparametric Additive Hazards Model (LS) and Survival Random Forest.

For our Neural Netwok we use one layer with 40 nodes, Elu activaction function with $\alpha = 0.1$, Adam optimizer, learning rate of 0.01, $\lambda = 0.09$ for Ridge penalty and 0.1 as Dropout rate. For DeepSurv we use the hyperparameters reported in their paper. 1 layer with 8 nodes, Selu activaction function, Adam optimizer, learning rate of 0.154, $\lambda = 6.551$ for Ridge penalty and 0.661 as Dropout rate. Moreover, in parenthesis we report the results reported by Katzman et al. (2018) for both DeepSurv and RSF. We plot in Figure 2.6b the ratio of the cumulative hazards between 3 groups defined by the 4 binary variables. It is clear from the plot how these ratios are not constant with time and therefore how the proportional hazards assumption, on which Deepsurv is based, is violated. From the results, our method indeed outperforms Deepsurv which is not showing better results than RSF. Moreover it outperforms RSF as well.

Table 2.13: Results for GBSG dataset

C-index	
Deep Hazard	**0.685**
Additive Hazards	0.666
Deep Surv	0.670 (0.676)
RSF	0.680 (0.648)

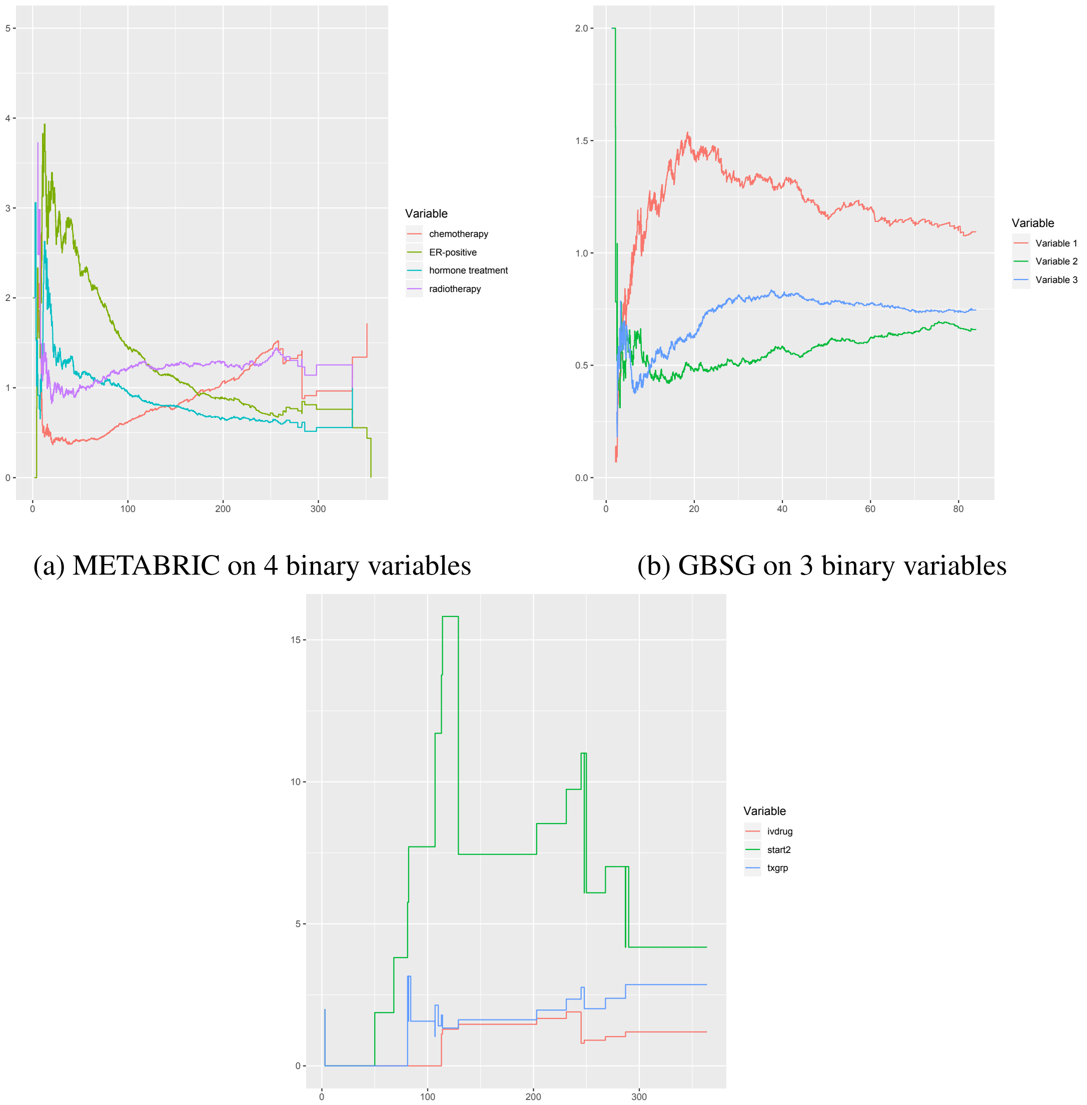
2.4.3 AIDS Clinical Trials Group (ACTG 320)

The dataset consists of 1151 HIV-infected patients (Hosmer et al., 2001). The data come from a double-blind, placebo-controlled trial that compared the three-drugs regime of indinavir, open label zidovudine (ZDV) or stavudine (d4T), and lamivudine (3TC) with the two-drugs regime of zidovudine or stavudine and lamivudine. Patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outcome measured was time to death and 2.26% of observations has observed death time. 500 observations are saved as test set.

We report in Table 2.14 the C-index for DeepHazard, DeepSurv, Semiparametric Additive Hazards Model (LS) and Survival Random Forest. For DeepHazard and DeepSurv we use 2 layers with 50 nodes, Selu activaction function, Adam optimizer, learning rate of 0.1, $\lambda = 2$ with Lasso penalty and 0.2 as Dropout rate. We plot in Figure 2.6c the ratio of the cumulative hazards between 3 groups defined by 3 binary variables ivdrug, start2 and txgrp, clearly indicating violation of proportionality of the hazards. We observe that our method outperforms DeepSurv and RSF.

Table 2.14: Results for AIDS:ACTG dataset

C-index	
Deep Hazard	**0.825**
Additive Hazards	0.824
Deep Surv	0.773
RSF	0.803



(a) METABRIC on 4 binary variables

(b) GBSG on 3 binary variables

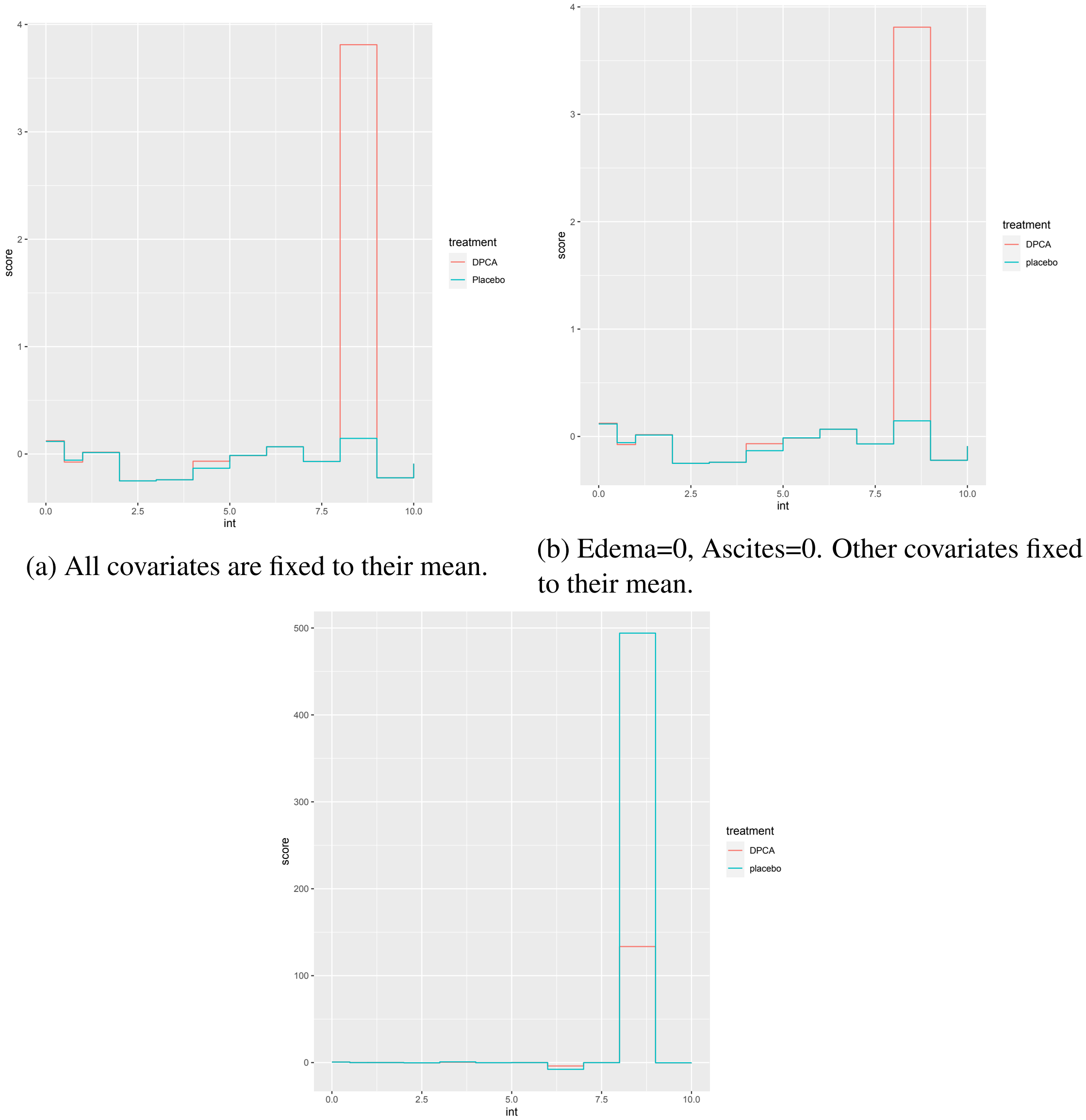
(c) AIDS on 3 binary variables

Figure 2.5: Proportional hazards diagnostic

2.4.4 Primary Biliary Cirrhosis: PBC dataset

We study the overall survival of patients with primary biliary cirrhosis, a fatal chronic liver disease.The popular PBC dataset comprise of 312 patients, referred to the Mayo Clinic between

January 1974 and May 1984, who participated in a randomized, double-blinded, placebo-controlled, clinical trial of the drug D-penicillamine. For each of the patients, clinical, biochemical, serologic and histologic parameters were collected. Since patients have been followed regularly since the trials ended, follow-up was extended to April 1988. By the end of the study, 140 of the 312 had died. The original clinical protocol for these patients specified visits at 6 months, 1 year, and annually thereafter. Fleming and Harrington (1991) studied the survival time of these patients using only the baseline value of the covariates. They showed that DPCA has a negligible effect on the survival, and they propose a model based on age, total serum bilirubin value, serum albumin value, prothrombin time and presence or absence of edema. This model, called the Mayo PBC model has been widely used and studied (Dickson et al., 1989; Jeffrey et al., 1990; Klion et al., 1992; Markus et al., 1989; Grambsch et al., 1989; Bonsel et al., 1990). Here, we propose to use Deephazard to analyse this dataset exploiting the available values of the covariates at each follow-up visits. To this aim, since we only have 3 deaths that happen after 11 years of being enrolled in the program, we consider the following set of time points $[0, 0.5, 1, 2, \ldots, 10, \infty]$. Moreover we assume that the value of the covariates is constant between one visit to the other, therefore, if a patient misses visit 3, we use as values of his/her covariates the ones collected at visit 2. Following the previous studies, we consider as covariates, the same one of the Mayo PBC model plus the presence of absence of ascites and the treatment. We include the presence of absence of ascites since Christensen et al. (1986) showed that it has a significant interaction with prednisone treatment and we use the treatment as covariate because we are interest in study the treatment effect on overall survival. We fit Deephazard with three layers with 10,15,10 nodes respectively. We use Elu(0.1) activation function, 0.2 of dropout, learning rate of 0.1, $\lambda = 10^{-4}$ with Ridge penalty, 1000 epochs and adam optimizer. Since in this case the majority of the covariates considered are internal, we don't predict the survival function but we predict for different hypothetical new patients the score $\hat{h}(Z(t), t)$. We predict the score $\hat{h}$ for two new patients with covariates at each time points fixed at their mean value and both treatment equal to placebo and DPCA . We plot the estimated score in figure 2.6. Moreover, we then predict the score $\hat{h}$ for someone with no edema or ascites both under treatment or placebo and for someone with both edema and ascites under treatment or placebo.We leave the other covariates fixed to their mean. Results are plotted in Figure 2.6.



(a) All covariates are fixed to their mean.

(b) Edema=0, Ascites=0. Other covariates fixed to their mean.

(c) Edema=1, Ascites=1. Other covariates fixed to their mean

Figure 2.6: PBC data: Predicted score for a new patient.

2.5 Discussion and possible applications

Although not extensively exploited in the past due to its complicated interpretation or the lack of methods available, understanding the relationship between survival and time-dependent covariates could be very useful in practice.

2.5.1 Individualized treatments

It could indeed be a helpful tool for making decisions in the context of dynamic treatment. Let's assume, for example, that, besides some baseline fixed covariates measured at the first visit, L_0, at each visit j, the doctor has to decide whether to put a patient under treatment, $A_j \in \{0,1\}$, which dose of certain medications to administer, $D_j \in [0,1]$, or whether continue with the same treatment or switch to some alternative, $M_j \in \{1,2,3\}$. The doctor could predict the survival of a new patient under different strategies and pick the one that maximizes patient survival. For example, at visit 2, considering the history of the covariate of a patient, $Z_{01} = (L_0, A_0.D_0, M_0, A_1, D_1, M_1)$, given two possible different strategies for visit 2, $z_2 = (a_2, d_2, m_2)$ and $z'_2 = (a'_2, d'_2, m'_2)$, the analysis of the predicted

$$\hat{S}(t \mid \{Z_{01}, z_2\}), \qquad \text{and} \qquad \hat{S}(t \mid \{Z_{01}, z'_2\}),$$

could help the doctor decides whether to treat the patient with strategy z_2 or z'_2. More in general, the same reasoning applies to other varying clinical variables as blood pressure. It could indeed be useful to observe the change in predicted survival under the different hypothetical paths of such covariates. If, for example, the increase of blood pressure appears dangerous for the patient, the doctor could think to introduce medications to keep it stable.

2.5.2 Estimation of treatment effects

Estimated conditional survival could also be needed as a necessary step towards obtaining a flexible estimator of some other parameter of interest. For example, it is common in the double robust treatment effect estimation literature to employ the use of method that require, as step one, the estimation of baseline quantity or conditional survival distribution. This is in particularly true when AIPW scores are constructed. The augmentation part of the latter indeed usually requires estimation of the conditional distribution of both the censoring and the time to event variable, (Zhang and Schaubel, 2012b; Zhao et al., 2015; Kang et al., 2018).

2.5.3 Variable predictive strength

On the other hand, the estimated conditional survival could be used to estimate other quantities of interest as the expected value of the survival time or R^2 measure of explained variation to study the predictive ability of different covariates. The latter is indeed function of the conditional variance of time T and it can be estimated, if an estimator $\hat{S}(t \mid Z(t))$ is available, using the following formula:

$$\widehat{\mathrm{Var}}\{T \mid Z(t)\} = \int_0^\tau 2t\hat{S}(t \mid Z(t))dt - \left\{\int_0^\tau \hat{S}(t \mid Z(t))dt\right\}^2.$$

Measure of explained variation can be used, for example, to evaluate the clinical importance of prognostic factors, the impact of genetic variants on gene expression on survival phenotypes or they can be applied in variable screening process, (Müller et al., 2008; Hielscher et al., 2010; Kong et al., 2019).

2.6 Appendix

2.6.1 Activation functions

Table 2.15: Activation functions

Atan	$a(x) = \mathrm{atan}(x)$
Elu(α)	$a(x) = \begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases}$
LeakyRelu	$a(x) = \begin{cases} x & x > 0 \\ 0.01x & x \leq 0 \end{cases}$
LogLog	$a(x) = 1 - \exp(-\exp(x))$
Relu	$a(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$
Selu	$a(x) = 1.0507 \begin{cases} x & x > 0 \\ 1.67326(e^x - 1) & x \leq 0 \end{cases}$
Tanh	$a(x) = \tanh(x)$

2.6.2 Technical details about $\bar{h}(t)$

We explain here why, for each j, $\bar{h}_{j,\theta_j}(t)$ is a step function with jump at censored event time X_i^j. We know that

$$\bar{h}_{j,\theta_j}(t) = \frac{\sum_{i=1}^{n_j} h_{j,\theta_j}(\tilde{Z}_i^j) Y_i^j(t)}{\sum_{i=1}^{n_j} Y_i^j(t)}, \tag{2.16}$$

and that, by definition,

$$Y_i^j(t) = \begin{cases} 1 & t \leq X_i^j \\ \\ 0 & t > X_i^j \end{cases} \tag{2.17}$$

Therefore, $\bar{h}_{j,\theta_j}(t)$ represents the mean of $h_{j,\theta_j}(\tilde{Z}^j)$ into the risk set at time t. Since the risk set changes only when an individual is censored or dies, $\bar{h}_{j,\theta_j}(t)$ changes only at censored event time X_i^j

2.6.3 Details on the estimation of cumulative hazard

If we break down everything we will have:

$$\begin{aligned} \lambda(t) &= \sum_{j=1}^{M+1} \lambda_0(t)\mathbb{1}(t_{j-1} < t \leq t_j) + \sum_{j=1}^{M+1} \mathbb{1}(t_{j-1} < t \leq t_j) h(Z_i(u), u) \\ &= \sum_{j=1}^{M+1} \lambda_0(t)\mathbb{1}(t_{j-1} < t \leq t_j) + \sum_{j=1}^{M+1} \mathbb{1}(t_{j-1} < t \leq t_j) h_j(\tilde{Z}_i^j), \end{aligned}$$

so if $\lambda_0^j(t) = \lambda_0(t)\mathbb{1}(t_{j-1} < t \leq t_j)$, we have:

$$\lambda(t) = \sum_{j=1}^{M+1} \left[\lambda_0^j(t) + h_j(\tilde{Z}_i^j)\right] \mathbb{1}(t_{j-1} < t \leq t_j).$$

Therefore, if we consider:

$$dN_i^j(t) = dM_i^j(t) + \int_{t_j}^t Y_i^j(u) d\Lambda(u|Z_i(u)), t \in [t_j, t_{j+1})$$

$$dN_i^j(t) = dM_i^j(t) + \int_{t_j}^t Y_i^j(u) d\Lambda^j(u|\tilde{Z}_i^j), t \in [t_j, t_{j+1}),$$

we have:

$$\hat{\Lambda}_0^j(t) = \begin{cases} 0 & t \leq t_j \\ \int_{t_j}^{t} \left[\sum_{i=1}^{n} Y_i^j(u)\right]^{-1} \sum_{i=1}^{n} \left(dN_i^j(u) - Y_i^j(u)\hat{h}_j(\tilde{Z}_i^j)\right) du & t_j < t \leq t_{j+1} \\ \int_{t_j}^{t_{j+1}} \left[\sum_{i=1}^{n} Y_i^j(u)\right]^{-1} \sum_{i=1}^{n} \left(dN_i^j(u) - Y_i^j(u)\hat{h}_j(\tilde{Z}_i^j)\right) du & t > t_{j+1} \end{cases}$$

and so:

$$\Lambda_0(t) = \sum_{j=1}^{J-1: t_{J-1}<t<t_J} \int_{t_{j-1}}^{t_j} \lambda_0(t) + \int_{t_{J-1}}^{t} \lambda_0(t) = \sum_{j=1}^{J: t_{J-1}<t<t_J} \int_{t_{j-1}}^{t_j} \lambda_0^j(t) + \int_{t_{J-1}}^{t} \lambda_0^J(t),$$

and so:

$$\hat{\Lambda}_0(t) = \sum_{j=1}^{M+1} \hat{\Lambda}_0^j(t),$$

and therefore:

$$\begin{aligned} \hat{\Lambda}_0(t) = & \sum_{j=1}^{M+1} [\int_{t_{j-1}}^{t_j} \left[\sum_{i=1}^{n} Y_i^j(u)\right]^{-1} \sum_{i=1}^{n} \left(dN_i^j(u) - Y_i^j(u)\hat{h}_j(\tilde{Z}_i^j)\right) du \mathbb{1}\{t > t_j\} \\ & + \int_{t_j}^{t} \left[\sum_{i=1}^{n} Y_i^j(u)\right]^{-1} \sum_{i=1}^{n} \left(dN_i^j(u) - Y_i^j(u)\hat{h}_j(\tilde{Z}_i^j)\right) du \mathbb{1}\{t_j < t < t_{j+1}\}] \\ = & \sum_{j=1}^{M+1} [\int_{t_{j-1}}^{t_j} \left[\sum_{i=1}^{n} Y_i(u)\right]^{-1} \sum_{i=1}^{n} \left(dN_i(u) - Y_i(u)\hat{h}(Z(u),u)\right) du \mathbb{1}\{t > t_j\} \\ & + \int_{t_j}^{t} \left[\sum_{i=1}^{n} Y_i(u)\right]^{-1} \sum_{i=1}^{n} \left(dN_i(u) - Y_i(u)\hat{h}(Z(u),u)\right) du \mathbb{1}\{t_j < t < t_{j+1}\}] \\ = & \int_{0}^{t} \frac{\sum_{i=1}^{n} \left\{dN_i(u) - Y_i(u)\hat{h}(Z_i(u),u)du\right\}}{\sum_{i=1}^{n} Y_i(u)}. \end{aligned}$$

2.7 Acknowledgements

Chapter 2, in full, is a reprint of the material as it may appear in Lifetime Data Analysis, 2021, Rava, Denise; Bradic, Jelena. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Doubly Robust Estimation of the Hazard Difference for Competing Risks Data

3.1 Introduction

Competing risks analysis concerns event times due to multiple causes. This work is motivated by a study on the effect of mid-life exposures on late-life cognitive outcomes related to Alzheimer's disease. We use data from the Honolulu Heart Program (HHP) and the Honolulu-Asia Aging Study (HAAS) on a cohort of Japanese men in Hawaii followed from 1965 to 2012 to investigate the effect of mid-life alcohol exposure on late-life cognitive impairment. As it is often the case in clinical trials, death is a competing risk for the event of interest; indeed, by the end of the study, only about 500 of the original 8006 men were still alive.

As for analysis of time-to-event data in general, it is often of interest to study the conditional treatment effect given the covariates in the presence of competing risks. Conditional treatment effects are typically expressed using regression models, and the commonly used ones include the proportional hazards model (Cox, 1972, 1975) and the additive hazards model (Aalen, 1980, 1989). The additive hazards model has received increasing attention lately because of its collapsibility, and therefore more suitable for causal inference (Tchetgen Tchetgen et al., 2015; Li et al., 2015; Zhao et al., 2015; Wang et al., 2017; Zheng et al., 2017; Ying et al., 2019). For a binary treatment, this conditional treatment effect is the hazard difference given the covariates under the additive hazards model. On the other hand, misspecification of the functional form of the covariates in the hazard regression model can lead to bias in the estimation of the treatment effect of interest.

To overcome such dependence on the correct specification of the covariate terms which are 'nuisance' themselves, flexible modeling such as nonparametric approaches might be considered. However, they are often inefficient and lead to slower rates of convergence of the estimated treatment effect; this is the 'curse of dimensionality' problem discussed in Robins and Ritov (1997). Alternatively, there has been a growing literature on doubly robust estimators that protect against misspecification of the 'nuisance' parts of the model (Robins and Rotnitzky, 1995, 2001; Bang and Robins, 2005; Tchetgen Tchetgen et al., 2010; Zhang and Schaubel, 2012a; Farrell, 2015; Jiang et al., 2017; Wang et al., 2017).

In the absence of competing risks, doubly robust estimators for the hazard difference have been proposed by Dukes et al. (2019b) and Hou et al. (2021). In the following we first derive the semiparametrically efficient score for the cause-specific hazard difference under competing risks. We then propose two doubly robust estimators with respect to two sets of models. The first set contains the treatment assignment model, also called the propensity score, and the model for the censoring distribution. The second set includes the cause-specific hazard models for the competing risks. As the proposed estimators incorporate the censoring distribution into the scores, they also weaken the assumption on censoring as needed in Hou et al. (2021) and Dukes et al. (2019b).

The rest of the paper is organized as follows: after formally defining the parameter of interest, in Section 3.3 we derive the two doubly robust scores. In Section 3.4 we describe their asymptotic properties and we derive their asymptotic distribution when the two sets of working models are both correct, or when only one of them is correct. We study the finite sample performance of the proposed estimators through extensive simulations in Section 3.5 and we then apply them on the HHP-HAAS dataset to estimate the effect of alcohol exposure on development of cognitive impairment in Section 3.6. We conclude with discussion in the last section.

3.1.1 Related work

In the context of time-to-event data, different doubly robust estimators have been proposed in the literature. The already mentioned works of Dukes et al. (2019b) and Hou et al. (2021) are most closely related to ours. They focus on doubly robust estimation of the constant difference between the hazard functions given the covariates in the absence of competing risks in low and high dimension, respectively.

Zhang and Schaubel (2012a); Bai et al. (2017); Sjölander and Vansteelandt (2017) derive doubly robust estimators for the treatment effect defined as the comparison between functions of the potential failure times T(1), T(0); i.e. the failure time that would be observed if a subject were treated or untreated, respectively. Zhang and Schaubel (2012a) and Bai et al. (2017) propose AIPW estimators for $E[f\{T(a)\}]$ for $a = 0,1$ and for different functions f. Sjölander and Vansteelandt (2017) develop a doubly robust estimator for the attributable fraction $1 - \frac{1-S_{T(0)}(t)}{1-S(t)}$. Yang et al. (2020) develop instead a doubly robust estimator for structural failure time models.

Another line of work discretizes the time,recasts the failure time as a 0-1 vector and uses techniques tailored for binary outcomes. For estimation of the parameters of marginal structural models, Petersen et al. (2014) and Zheng et al. (2016) derive targeted maximum likelihood estimators that are doubly robust while Yu and Van Der Laan (2006) propose a doubly robust estimator following Van Der Laan et al. (2003) theory.

3.1.2 Model and Notation

Assume there are J competing risks and denote $T_1, \ldots, T_J$ the (latent) time to each type of failure. Let $T = \min(T_1, \ldots, T_J)$, C be the censoring random variable, and $X = \min(T, C)$ be the observed (and possibly censored) failure time. Denote $\delta = \mathbb{1}\{T \leq C\}$ the event indicator, and let $\varepsilon = 1, \ldots, J$ indicate the type of failure. Let $A = 0$, 1 be a binary treatment, and Z be a vector of baseline covariates.

A commonly used approach for competing risks data is to model the cause-specific hazard function for each type of failure (Holt, 1978; Benichou and Gail, 1990; Kalbfleisch and Prentice, 2011). The cause-specific hazard functions are the quantities 'just identified' by such data, in the sense that any other quantity that can be identified from competing risks data, can be expressed as a function of the cause-specific hazard (Kalbfleisch and Prentice, 2011). We assume that the conditional cause-specific hazard function, $h_j(t|A,Z) = \lim_{\Delta_t \to 0} \frac{1}{\Delta_t} P(t \leq T < t + \Delta_t, \varepsilon = j | T \geq t, A, Z)$, for $j = 1, \ldots, J$, satisfies:

$$h_j(t|A,Z) = \beta_j A + \lambda_j(t,Z), \tag{3.1}$$

where $\lambda_j(t,Z)$, representing the effect of the covariates on the hazard, is left unspecified. This is a key difference from the more traditional cause-specific additive hazards model that assumes linear effects of both A and Z; see for example, Shen and Cheng (1999). From model (3.1) then, $\beta_j = h_j(t|A = 1, Z) - h_j(t|A = 0, Z)$ is the difference between the conditional cause-specific hazard functions of the two treatment groups.

In the following we assume that $C \perp T | (A, Z)$, where '$\perp$' indicates statistical independence.

This is a standard assumption in the analysis of time-to-event data, and it relaxes the stricter assumption $C \perp (A,T)|Z$ imposed by both Hou et al. (2021) and Dukes et al. (2019b). We will also use the counting process and the at-risk process notation: $N_j(t) = \mathbb{1}\{X \leq t,\ \delta = 1,\ \varepsilon = j\}$ and $Y(t) = \mathbb{1}\{X \geq t\}$. Under model (3.1), $M_j(t) = N_j(t) - H_j(t|A,Z)Y(t)$ is a local square-integrable martingale with respect to the filtration $\mathcal{F}_t = \sigma\left\{N_j(s), Y(s+), A, Z : j = 1,\ldots,J,\ 0 < s < t\right\}$, where $H_j(t|A,Z) = \int_0^t h_j(u|A,Z)du$.

3.2 Semiparametrically efficient score for β

In the following we derive the orthogonal complement of the nuisance tangent space and the efficient score for $\beta = [\beta_1,\ldots,\beta_J]^\top$. The derivation follows the modern semiparametric theory as described in Tsiatis (2007), and we provide the details in Section 3.8.1 of the Supplement.

Under model 3.1, the data follows a semiparametric distribution identified by the parameter of interest $\beta = [\beta_1,\ldots,\beta_J]^\top$ and the nuisance parameter $\eta = [\lambda_1(t,z),\ldots,\lambda_J(t,z),\lambda_c(t|a,z)$ $,p(a|z),f(z)]^\top$, where $\lambda_c(t|a,z)$ is the conditional hazard function for C, $P(a|z)$ is the conditional distribution of A and $f(z)$ is the density of the covariates. The likelihood for a single copy of the data takes indeed the following form:

$$\begin{aligned} L &= \prod_{j=1}^{J} \left\{\beta_j A + \lambda_j(X,Z)\right\}^{\mathbb{1}\{\delta=1,\varepsilon=j\}} \exp\left\{-\beta_j A t - \Lambda_j(X,Z)\right\} \\ &\quad \times \{\lambda_c(X|A,Z)\}^{1-\delta} \exp\{-\Lambda_c(X|A,Z)\}\, p(A|Z) f(Z), \end{aligned}$$

where $\Lambda_j(t,z) = \int_0^t \lambda_j(u,z)du$ for $j = 1,\ldots,J$ and $\Lambda_c(t|a,z) = \int_0^t \lambda_c(u|a,z)du$. From the likelihood, one can derive the score for the parameter of interest, $S_\beta = \frac{\partial \log L}{\partial \beta}$ and, if η has finite dimension, the score for the nuisance parameter, $S_\eta = \frac{\partial \log L}{\partial \eta}$. In this case, the nuisance tangent space is the space spanned by the nuisance score. When η has infinite dimension, as in our case, the notion of nuisance tangent space can be extended through the definition of parametric submodels. We leave the technicality of this definition to Chapter 4 of Tsiatis (2007).

An estimator $\hat{\beta}$ is asymptotically linear if there exists a function of the data φ, such that $\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_i + o_p(1)$. The function φ, named influence function, has mean zero and finite variance and guarantees the asymptotic normality of the estimator. Such estimators are therefore desirable and they are uniquely defined by their influence function. Theorem 4.2. of Tsiatis (2007) proves that every influence function belongs to the orthogonal complement of the nuisance tangent space. This space, denoted by $\Lambda^{\perp}$, is therefore the starting point to define semiparametric estimators for β that are consistent and asymptotically normal.

The space $\Lambda^{\perp}$ is also important since it allows one to find orthogonal scores in the classical sense of the definition. A score $\psi(\beta, \eta)$ is orthogonal if

$$\frac{\partial}{\partial r} \mathrm{E}\{\psi(\beta_0; \eta_0 + r(\eta - \eta_0))\}\bigg|_{r=0} = 0,$$

where we use the subscript 0 to indicate the true parameters. Orthogonal scores are invariant to small perturbations of the nuisance parameter around the true and so the estimation of the nuisance parameter doesn't greatly affect the estimation of the treatment effect (Bickel et al., 1993; Newey, 1990, 1994). Lemma 11 in the Supplement shows that an estimating function belongs to $\Lambda^{\perp}$ if and only if it is orthogonal.

The following lemma, proven in the Supplement, defines the form of the orthogonal complement of the nuisance tangent space.

Lemma 1. *Under model* (3.1)*, the orthogonal complement of the nuisance tangent space takes the* *following form:*

$$\begin{aligned}\Lambda^{\perp} &= \Bigg\{\sum_{j=1}^{J}\int_0^{\tau}\left[g_j(t,A,Z)-\frac{E\left\{g_j(t,A,Z)h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}{E\left\{h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}\right]\\ &\quad\times\frac{dM_j(t)}{h_j(t|A,Z)}:\ \textit{for all}\ g_j(t,A,Z)\in\mathbb{R}^J\Bigg\}.\end{aligned}\tag{3.2}$$

Among all the semiparametric asymptotically linear estimators of β it is often of interest to derive the efficient one. Tsiatis (2007) defines the efficient score as $S_\beta-\Pi\{S_\beta|\Lambda\}$. Since, under model (3.1), $S_\beta=\left\{\int_0^\tau A\frac{dM_j(t)}{h_j(t|A,Z)}\right\}_{j=1}^{J}$, we have the following lemma.

Lemma 2. *Under model* (3.1) *the efficient score has the following form:*

$$S_{eff} = \left\{\int_0^{\tau}\left[A-\frac{E\left\{Ah_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}{E\left\{h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}\right]\frac{dM_j(t)}{h_j(t|A,Z)}\right\}_{j=1}^{J}. \tag{3.3}$$

The above score is locally efficient in the sense that its asymptotic variance attains the semiparametric efficency bound when $P(a|z), S_c(t|a,z)$ and $\lambda_j(t,z)$ are known or correctly estimated (Theorem 4.1. of Tsiatis (2007)). Unfortunately, since $h_j(t|A,Z)$ in (3.3) is unknown and estimators for it are not readily available, the efficient score may not be directly used in practice. We will however exploit both (3.2) and (3.3) to derive two doubly robust scores for estimation of β.

Remark: if we make the stronger assumption of $C\perp(A,T)|Z$ as in Dukes et al. (2019b) and Hou et al. (2021), $S_c(t|A,Z)=S_c(t|Z)$, and so the efficient score simplifies to:

$$S_{eff} = \left\{\int_0^{\tau}\left[A-\frac{E\left\{Ah_j^{-1}(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}{E\left\{h_j^{-1}(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}\right]\frac{dM_j(t)}{h_j(t|A,Z)}\right\}_{j=1}^{J}.$$

In this case, S_c is therefore no longer needed for the estimation of β.

3.3 Doubly robust scores

Doubly robust score 1: Inspired by Hou et al. (2021), we choose $g_j(t,A,Z) = \{A - \pi(Z)\} h_j(t|A,Z) S_c^{-1}(t|A,Z) e^{\sum_{l=1}^J \beta_l A t}$ in (3.2). We obtain the following estimating function:

$$S_1(\beta; A, Z, S_c, \pi, \Lambda) = \left\{ \int_0^\tau e^{\sum_{l=1}^J \beta_l A t} S_c^{-1}(t|A,Z) \{A - \pi(Z)\} \, dM_j(t;\beta,\Lambda) \right\}_{j=1}^J . \qquad (3.4)$$

Here we have used the propensity score notation $\pi(Z) = P(A=1|Z)$ and $M_j(t;\beta,\Lambda) = N_j(t) - Y(t)\beta_j A t - Y(t)\Lambda_j(t,Z)$. We use $h_j(t|A,Z)$ in the definition of g_j to cancel the hazard weights $h_j^{-1}(t|A,Z)$ from (3.2). To understand the rest of g_j is important to notice that, under model (3.1), $\mathrm{E}\{Y(t)|A,Z\} = e^{-\sum_{j=1}^J \beta_{j0} A t} S_{c0}(t|A,Z) e^{-\sum_{j=1}^J \Lambda_{j0}(t,Z)}$, where again the subscript 0 is used to indicate the true quantities. The expectation of the $j^{th}-$component of S_1 with the true parameter of interest plugged in is:

$$\begin{aligned} &\mathrm{E}\left(\int_0^\tau \mathrm{E}\left[e^{\sum_{l=1}^J \beta_{l0} A t} S_c^{-1}(t|A,Z) \{A - \pi(Z)\} \mathrm{E}\{Y(t)|A,Z\} |Z \right] d\left\{ \Lambda_j(t,Z) - \Lambda_{j0}(t,Z) \right\} \right) \\ = &\mathrm{E}\left(\int_0^\tau \mathrm{E}\left[S_c^{-1}(t|A,Z) S_{c0}^{-1}(t|A,Z) \{A - \pi(Z)\} |Z \right] e^{-\sum_{l=1}^J \Lambda_{l0}(t,Z)} d\left\{ \Lambda_j(t,Z) - \Lambda_{j0}(t,Z) \right\} \right). \end{aligned}$$

Therefore, the form of g_j is chosen such that, as it is common for doubly robust scores, the above integrand is the product of two residuals, one for the outcome models for the competing risks and one for the censoring and the treatment model.

The main difference between our score and Hou et al. (2021) score is that, beside our score being suitable for estimation in a competing risks setting, our score, incorporating the censoring distribution S_c, does not need the stronger assumption $C \perp (T,A)|Z$ to hold.

Given quantities $S_c(\cdot|\cdot,\cdot)$, $\pi(\cdot)$ and $\Lambda(\cdot,\cdot) = [\Lambda_1(\cdot,\cdot), \ldots, \Lambda_J(\cdot,\cdot)]^\top$ we propose the following score for estimation of β:

$$S_{1,n}(\beta; S_c, \pi, \Lambda) := \frac{1}{n}\sum_{i=1}^{n} S_1(\beta; A_i, Z_i, S_c, \pi, \Lambda) = 0. \tag{3.5}$$

Doubly robust score 2: Traditionally, for hazard models of the additive form, the hazard weights have been removed from the efficient score to derive scores that can be used in practice (Lin and Ying, 1994b). If we simplify the efficient score (3.3), removing the hazard weights, we are left with the following:

$$S_2(\beta; A, Z, S_c, \pi, \Lambda) \quad = \left\{ \int_0^\tau \left\{ A - \mathcal{E}_A(t; \beta, S_c, \pi, Z) \right\} dM_j(t; \beta, \Lambda) \right\}_{j=1}^{J}, \tag{3.6}$$

where:

$$\begin{aligned} \mathcal{E}_A(t; \beta, S_c, \pi, Z) &= \frac{\mathrm{E}\left[A e^{-\sum_{j=1}^{J} \beta_j A t} S_c(t|A, Z) | Z \right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{J} \beta_j A t} S_c(t|A, Z) | Z \right]} \\ &= \frac{e^{-\sum_{j=1}^{J} \beta_j t} S_c(t|A=1, Z)\pi(Z)}{e^{-\sum_{j=1}^{J} \beta_j t} S_c(t|A=1, Z)\pi(Z) + S_c(t|A=0, Z)\left\{1 - \pi(Z)\right\}}. \end{aligned}$$

Given quantities $S_c(\cdot|\cdot,\cdot)$, $\pi(\cdot)$ and $\Lambda(\cdot,\cdot) = [\Lambda_1(\cdot,\cdot), \ldots, \Lambda_J(\cdot,\cdot)]^\top$ we propose the following score for estimation of β:

$$S_{2,n}(\beta; S_c, \pi, \Lambda) := \frac{1}{n}\sum_{i=1}^{n} S_2(\beta; A_i, Z_i, S_c, \pi, \Lambda) = 0. \tag{3.7}$$

Since the two proposed scores belong to $\Lambda^\perp$, they are orthogonal. Moreover they are doubly robust with respect to the estimation of both $S_c(\cdot|\cdot,\cdot)$ and $\pi(\cdot)$ and of $\Lambda(\cdot,\cdot)$.

Theorem 2. *$E\{S_1(\beta_0; A, Z, S_c, \pi, \Lambda)\} = 0$ and $E\{S_2(\beta_0; A, Z, S_c, \pi, \Lambda)\} = 0$ if either*

$\{S_c(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi(\cdot) = \pi_0(\cdot)\}$ or $\Lambda(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$, where we use subscript 0 to indicate the true quantities.

Score 2, (3.6), is completely new, no similar score has never been proposed in the literature, even in the absence of competing risks.

Both scores incorporate the censoring distribution, relaxing the censoring assumption of both Hou et al. (2021) and Dukes et al. (2019b). However, if we are willing to make the stronger assumption $C \perp (T,A)|Z$, the two scores simplify to:

$$\begin{aligned} \tilde{S}_1(\beta;A,Z,\pi,\Lambda) &= \left\{\int_0^\tau e^{\sum_{j=1}^J \beta_j A t}\{A-\pi(Z)\}\, dM_j(t;\beta,\Lambda)\right\}_{j=1}^J, & (3.8)\\ \tilde{S}_2(\beta;A,Z,\pi,\Lambda) &= \left\{\int_0^\tau \left\{A - \frac{e^{-\sum_{j=1}^J \beta_j t}\pi(Z)}{e^{-\sum_{j=1}^J \beta_j t}\pi(Z)+\{1-\pi(Z)\}}\right\} dM_j(t;\beta,\Lambda)\right\}_{j=1}^J. & (3.9)\end{aligned}$$

Traditionally, estimation of parameters in competing risks setting has been carried over estimating one parameter at a time and considering the other competing risks as censoring. This is true for example for the cause-specific Cox proportional hazards model and the traditional cause-specific additive hazards model. A novelty of both our approaches is that every component of our parameter of interest β is here estimated together using a multidimensional score.

3.4 Estimation and inference

Both proposed scores depend on the quantities $S_c(\cdot|\cdot,\cdot), \pi(\cdot), \Lambda(\cdot,\cdot)$. These, unknown in observational studies, can be estimated via working models. Once estimators $\hat{S}_c(\cdot|\cdot,\cdot), \hat{\pi}(\cdot), \hat{\Lambda}(\cdot,\cdot)$ are available, we define $\hat{\beta}^{(1)}$ to be the root of $S_{1,n}(\beta;\hat{S}_c,\hat{\pi},\hat{\Lambda})$ and $\hat{\beta}^{(2)}$ to be the root of $S_{2,n}(\beta;\hat{S}_c,\hat{\pi},\hat{\Lambda})$. The user is free to choose any working model as long as mild traditional assumptions, listed later, are satisfied. For estimation of the propensity score $\pi(\cdot)$ and the censoring model $S_c(\cdot|\cdot,\cdot)$ we don't offer any specific suggestion. Estimation of $\Lambda(\cdot,\cdot)$ is more delicate; we propose to use the following set of linear working models: $\left\{\Lambda_j(t,Z;G_j,\gamma_j) = G_j(t) + \gamma_j^\top Zt\right\}_{j=1}^{J}$. The parameters $\gamma = [\gamma_1,\ldots,\gamma_J]^\top$ and $G = [G_1,\ldots,G_J]^\top$ can be estimated applying the cause-specific additive hazards model routine (Shen and Cheng, 1999). This applies the estimating procedures proposed by Lin and Ying (1994b) separately to each competing risk:

$$\hat{\gamma}_j = \left[\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{Z_i - \bar{Z}(t)\}^{\otimes 2}dt\right]^{-1}\left[\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{Z_i - \bar{Z}(t)\}dN_{ji}(t)\right], \tag{3.10}$$

where $\bar{Z}(t) = \{\sum_{i=1}^n Y_i(t)\}^{-1}\sum_{i=1}^n Y_i(t)Z_i$, $Z^{\otimes 2} = Z^\top Z$ and

$$\hat{G}_j(t;\beta_j,\gamma_j) = \int_0^t \frac{\sum_{i=1}^n \left\{dN_{ji}(u) - Y_i(u)\beta_j A_i du - Y_i(u)\gamma_j^\top Z_i du\right\}}{\sum_{i=1}^n Y_i(u)}. \tag{3.11}$$

For estimation of G_j we moreover propose the following weighted version of the Breslow estimator:

$$\hat{G}_j(t;\beta_j,\gamma_j,S_c,\pi) = \int_0^t \frac{\sum_{i=1}^n w_i(S_c,\pi)\left\{dN_{ji}(u) - Y_i(u)\beta_j A_i du - Y_i(u)\gamma_j^\top Z_i du\right\}}{\sum_{i=1}^n w_i(S_c,\pi)Y_i(u)}, \tag{3.12}$$

where $w_i(S_c,\pi) = S_c^{-1}(u|A_i,Z_i)A_i\{1-\pi(Z_i)\}$. The above weights are chosen such that plugging (3.12) as estimator of $G_j(t)$ into (3.5) gives the following closed form score:

$$\begin{aligned} & S_{1,n}(\beta;\hat{S}_c,\hat{\pi},\hat{\Lambda}) & (3.13)\\ = \quad & \Bigg\{-\beta_j\frac{1}{n}\sum_{i=1}^n\int_0^\tau \hat{S}_c^{-1}(t|A_i,Z_i)(1-A_i)\hat{\pi}(Z_i)Y_i(t)dt - \frac{1}{n}\sum_{i=1}^n\int_0^\tau \hat{S}_c^{-1}(t|A_i,Z_i)(1-A_i)\hat{\pi}(Z_i) \\ & \cdot\left(dN_{ji}(t) - Y_i(t)\left[\hat{\gamma}_j^\top\left\{Z_i - \bar{Z}(t;\hat{S}_c,\hat{\pi})\right\}dt + d\bar{N}_j(t;\hat{S}_c,\hat{\pi})\right]\right)\Bigg\}_{j=1}^{J}, \end{aligned}$$

where:

$$\bar{Z}(t;\hat{S}_c,\hat{\pi}) = \frac{\sum_{i=1}^n Y_i(t) Z_i w_i(\hat{S}_c,\hat{\pi})}{\sum_{i=1}^n Y_i(t) w_i(\hat{S}_c,\hat{\pi})}, \quad d\bar{N}_j(t;\hat{S}_c,\hat{\pi}) = \frac{\sum_{i=1}^n dN_{ji}(t) w_i(\hat{S}_c,\hat{\pi})}{\sum_{i=1}^n Y_i(t) w_i(\hat{S}_c,\hat{\pi})}.$$

One could argue that using only the treated subjects for estimation of G could lead to a loss in the efficiency of the estimator of the parameter of interest. However, we show in the next section that, when all the nuisance parameters are consistently estimated, the asymptotic distribution of $\hat{\beta}^{(1)}$ does not depend on the specific estimator of $\Lambda(\cdot)$.

No weighted version of the Breslow estimator can lead to a closed-form expression for $\hat{\beta}^{(2)}$. Plugging (3.11) in (3.6), after some tedious algebra, we get the following score:

$$\begin{aligned} S_{2,n}(\beta;\hat{S}_c,\hat{\pi},\hat{\Lambda}) \quad = \quad & \left\{ \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ A_i - \mathcal{E}_{A_i}(t;\beta,\hat{S}_c,\hat{\pi},Z_i) - \bar{A}(t) + \bar{\mathcal{E}}(t) \right\} \right. \\ & \left. \cdot \left\{ dN_{ji}(t) - Y_i(t) \left(\beta_j A_i + \hat{\gamma}_j^\top Z_i \right) dt \right\} \right\}_{j=1}^J, \end{aligned} \tag{3.14}$$

where

$$\bar{A}(t) = \frac{\sum_{i=1}^n Y_i(t) A_i}{\sum_{i=1}^n Y_i(t)}, \quad \bar{\mathcal{E}}(t) = \frac{\sum_{i=1}^n Y_i(t) \mathcal{E}_{A_i}(t;\beta,\hat{S}_c,\hat{\pi},Z_i)}{\sum_{i=1}^n Y_i(t)}.$$

Since however we leave to the user the freedom to choose any working models, from now on, we use $\hat{S}_c(\cdot|\cdot,\cdot),\hat{\pi}(\cdot),\hat{\Lambda}(\cdot,\cdot)$ to denote generic estimators for the nuisance parameters $S_c(\cdot|\cdot,\cdot),\pi(\cdot),\Lambda(\cdot,\cdot)$. It is possible that the estimator used for $\Lambda(\cdot,\cdot)$ depends on β, as the proposed (3.11) and (3.12); we will therefore use as generic notation $\hat{\Lambda}(t,z;\beta)$ when we leave the estimator $\hat{\Lambda}(\cdot,\cdot)$ to depend on the unknown parameter.

We will now study the asymptotic properties of score 1 and score 2. For ease of notation we report the results assuming $J = 2$; however, everything can be extended to the case of more than 2 competing risks.

3.4.1 Asymptotic properties of score 1.

To prove consistency and asymptotic normality of $\hat{\beta}^{(1)}$ we need a series of assumptions.

Assumption 1. *There exist $S_c^*(\cdot|\cdot,\cdot), \pi^*(\cdot), \Lambda^*(\cdot,\cdot)$ such that:*

$$\begin{aligned}
\sup_{t\in[0,\tau],z\in\mathcal{Z},a=0,1} \left|\hat{S}_c(t|a,z) - S_c^*(t|a,z)\right| &= O_p(a_n),\\
\sup_{z\in\mathcal{Z}} |\hat{\pi}(z) - \pi^*(z)| &= O_p(b_n),\\
\sup_{t\in[0,\tau],z\in\mathcal{Z}} \left|\hat{\Lambda}_j(t,z;\beta_{j0}) - \Lambda_j^*(t,z)\right| &= O_p(c_n),
\end{aligned}$$

for some $a_n = o(1)$, $b_n = o(1)$, $c_n = o(1)$ and for $j = 1,2$.

Assumption 1, common for the literature on doubly robust estimators (Zhang and Schaubel, 2012a; Yang et al., 2020), assumes that the generic estimators $\hat{S}_c(\cdot|\cdot,\cdot), \hat{\pi}(\cdot), \hat{\Lambda}(\cdot,\cdot)$ converge to some $S_c^*(\cdot|\cdot,\cdot), \pi^*(\cdot), \Lambda^*(\cdot,\cdot)$, possibly different from the true quantities. We don't require specific rates of convergence for the estimators of the nuisance parameters; assumptions on a_n, b_n, c_n will indeed depend on which model is correctly specified. We moreover need a series of common regularity assumptions that we report in Section 3.8.3 of the Supplement.

The following results prove consistency of our estimator as long as one of the sets of models is correctly specified. Moreover they prove our estimator to be both rate-doubly robust and model-doubly robust in the sense that it is asymptotically normal if either both sets of models are correctly specified and the product of their convergence rates is $o(n^{-1/2})$ or if only one of the two sets of models is correctly specified with a convergence rate of $\sqrt{n}$.

Theorem 3. *Let Assumption 1 and Assumptions S1-S8 in the Supplement hold. If either* $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ *and* $\pi^*(\cdot) = \pi_0(\cdot)$ *or* $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$*, it holds* $\hat{\beta}^{(1)} - \beta_0 = o_p(1)$.

Theorem 4. *Let Assumption 1 and Assumptions S1-S8 in the Supplement hold.*

a) (Model-double robustness): Let $S_c^*(t|a,z) = S_c(t|a,z;\eta_0,\Lambda_{c0}) = S_{c0}(t|a,z)$, $\pi^*(z) = \pi(z;\alpha_0) = \pi_0(z)$ *and* $\Lambda^*(\cdot,\cdot) \neq \Lambda_0(\cdot,\cdot)$*, for some known functions* S_c *and* π*. Let* $a_n = b_n = n^{-1/2}$*; specifically let Assumptions A1-A2 in the Supplement hold. Then* $\sqrt{n}\left(\hat{\beta}^{(1)} - \beta_0\right)$ *is asymptotically linear with influence function* $\left\{K^{(a)}\right\}^{-1}\psi^{(a)}$ *and therefore,*

$$\sqrt{n}\left(\hat{\beta}^{(1)} - \beta_0\right) \xrightarrow{D} \mathcal{N}(0, \Sigma^{(a)} = \left\{E(K^{(a)})^{-1}\right\}^{\top} Var(\psi^{(a)}) E(K^{(a)})^{-1}).$$

b) (Model-double robustness): Let $\Lambda^*(t,z) = L(t,z;G_0,\gamma_0) = \Lambda_0(t,z)$, $S_c^*(\cdot|\cdot,\cdot) \neq S_{c0}(\cdot|\cdot,\cdot)$ *and* $\pi^*(\cdot) \neq \pi_0(\cdot)$*, for some known function* L*. Let* $c_n = n^{-1/2}$*; specifically let Assumptions B1-B2 in the Supplement hold. Then* $\sqrt{n}\left(\hat{\beta}^{(1)} - \beta_0\right)$ *is asymptotically linear with influence function* $\left\{K^{(b)}\right\}^{-1}\psi^{(b)}$ *and therefore,*

$$\sqrt{n}\left(\hat{\beta}^{(1)} - \beta_0\right) \xrightarrow{D} \mathcal{N}(0, \Sigma^{(b)} = \left\{E(K^{(b)})^{-1}\right\}^{\top} Var(\psi^{(b)}) E(K^{(b)})^{-1}).$$

c) (Rate-double robustness): If $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot) = \pi_0(\cdot)$ *and* $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ *with* $a_n c_n = o(n^{-1/2})$ *and* $b_n c_n = o(n^{-1/2})$ *under Assumptions C1, C2 in the the Supplement*

$$\sqrt{n}(\hat{\beta}^{(1)} - \beta_0) \xrightarrow{D} \mathcal{N}(0, \Sigma^{(c)} = \left\{(W^{(c)})^{-1}\right\}^{\top} V^{(c)} (W^{(c)})^{-1}),$$

where $V^{(c)}$ *and* $W^{(c)}$ *are diagonal matrices with components* $\int_0^\tau \left\{p(u)\beta_{j0} + q_j(u)\right\} du$ *and* $E\left(\int_0^\tau e^{(\beta_{10}+\beta_{20})At} A \left\{S_{c0}(t|A,Z)\right\}^{-1} \left\{A - \pi_0(Z)\right\} Y(t) dt\right)$*, respectively.*

Quantities $\psi^{(a)}, \psi^{(b)}$ and $K^{(a)}, K^{(b)}$ are given in Section 3.8.4 of the Supplement. Quantities p and q are defined in Assumption C1 in the Supplement.

Remark 1. *Both the assumed working models for $S_c(\cdot|\cdot,\cdot)$ and $\Lambda(\cdot,\cdot)$ are semiparametric. They indeed have a parametric component encoded by η and γ and a nonparametric component encoded by $\Lambda_c(t)$ and $G(t)$, respectively. Either one of them can be chosen null by the user.*

The consistency result requires either sets of estimators of the nuisance parameters to converge to the true without requiring any specific rate of convergence or knowledge of which model is correct. The asymptotic normality of the score requires more specific assumptions. Case a) and b) of Theorem 4 assume that only one of the two sets of models is correctly specified and that the rate of convergence of the corresponding estimators is $\sqrt{n}$. This is easily achieved using classical semiparametric models as logistic regression for the propensity score and the Cox model for the censoring distribution. Both our proposals for estimation of $\Lambda(t,Z)$ achieve the required rate of convergence. Case a) and b) of of Theorem 4 provide the asymptotic distribution of $\hat{\beta}^{(1)}$ when one of the two sets of models is possibly misspecified. This is an improvement with respect to the result of Hou et al. (2021), where the asymptotic distribution of the estimator is derived only when both models are correct. Case c) of Theorem 4 assumes that both sets of models are correctly specified. If this is the case no specific rate is required for the convergence of the estimators of the nuisance parameters, as long as their product rate is $o(n^{-1/2})$. A set of estimators can therefore be arbitrary slow as long as the other set is fast enough. The property just described is known as rate double robustness and allows the user to choose from a variety of estimators for the nuisance parameters. This is a relaxation with respect to Wang and Chen (2001); Tchetgen Tchetgen et al. (2010); Zhang and Schaubel (2012a); Bai et al. (2017); Dukes et al. (2019b); Tan (2019). The following result derives a consistent estimator for the asymptotic variance of $\hat{\beta}^{(1)}$ when both models are correctly specified.

Theorem 5. *Let Assumption 1 and Assumptions S1-S8 in the Supplement hold. If* $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$*,* $\pi^*(\cdot) = \pi_0(\cdot)$ *and* $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ *with* $a_n c_n = o(n^{-1/2})$ *and* $b_n c_n = o(n^{-1/2})$ *under Assumptions C1, C2 in the the Supplement, the asymptotic variance of* $\hat{\beta}^{(1)}$ *can be consistently estimated by:*

$$\left\{\hat{W}^{(c)}\right\}^{-1} \hat{V}^{(c)}(\tau) \left\{\hat{W}^{(c)}\right\}^{-1}, \tag{3.15}$$

where

$$\hat{W}_{jj}^{(c)} = \frac{1}{n}\sum_{i=1}^{n} A_i \left\{A_i - \hat{\pi}(Z_i)\right\} \int_0^{X_i} \left\{\hat{S}_c(t|A_i,Z_i)\right\}^{-1} e^{(\hat{\beta}_1^{(1)}+\hat{\beta}_2^{(1)})t} dt,$$

and

$$\hat{V}_{jj}^{(c)}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\delta_i = 1, \varepsilon_i = j\} e^{2(\hat{\beta}_1^{(1)}+\hat{\beta}_2^{(1)})A_i X_i} \hat{S}_c^{-2}(X_i|A_i,Z_i) \left\{A_i - \hat{\pi}(Z_i)\right\}^2,$$

for $j = 1,2$*.*

The asymptotic variance when one of the two sets of models is misspecified, (case a) and b) of Theorem 4), depends on the form of the nuisance parameters correctly specified and their estimators. In the following corollary of case a) and b) of Theorem 4 we derive the explicit form of the asymptotic variance for common specific working models.

Corollary 1. *Under the same Assumptions of Theorem 4, it holds:*

a) If $S_c^*(t|a,z) = S_{c0}(t|a,z) = \exp\left(-\Lambda_{c0}(t)e^{\eta_0^\top d}\right)$ *where* $D = [A,Z]'$*,* $\pi^*(z) = \pi_0(z) = \left\{1 + \exp(-\alpha^\top z)\right\}^{-1}$ *and* $\Lambda^*(\cdot,\cdot) \neq \Lambda_0(\cdot,\cdot)$*, and if additional regularity Assumptions A*1-4 in the*

Supplement hold, then

$$\sqrt{n}\left(\hat{\beta}^{(1)}-\beta_1\right)\xrightarrow{D}\mathcal{N}(0,\Sigma^{(a')}=\left\{E(K^{(a)})^{-1}\right\}^{\top}Var(\psi^{(a')})E(K^{(a)})^{-1}).$$

b) Let' s assume that $\Lambda^*(t,z)=\Lambda_0(t,z)=G_0(t)+\gamma_0^{\top}zt$ *,* $S_c^*(\cdot|\cdot,\cdot)\neq S_{c0}(\cdot|\cdot,\cdot)$ *and* $\pi^*(\cdot)\neq\pi^o(\cdot)$.

If $G_j(t)$ *is estimated using* (3.11) *then, under additional regularity Assumptions B*1 ,2, 4 in the Supplement,*

$$\sqrt{n}\left(\hat{\beta}^{(1)}-\beta_0\right)\xrightarrow{D}\mathcal{N}(0,\Sigma^{(b')}=\left\{E(K^{(b')})^{-1}\right\}^{\top}Var(\psi^{(b')})E(K^{(b')})^{-1}).$$

If $G_j(t)$ *is estimated using* (3.12) *then,under additional regularity Assumptions B*1 ,3, 4 in the Supplement, then*

$$\sqrt{n}\left(\hat{\beta}^{(1)}-\beta_0\right)\xrightarrow{D}\mathcal{N}(0,\Sigma^{(b'')}=\left\{E(K^{(b'')})^{-1}\right\}^{\top}Var(\psi^{(b'')})E(K^{(b'')})^{-1}).$$

Quantities $K^{(a)},K^{(b')},K^{(b'')},\psi^{(a')},\psi^{(b')},\psi^{(b'')}$ *are given in Section 3.8.4 of the Supplement.*

The above Corollary offers an explicit form for the asymptotic variance of $\hat{\beta}^{(1)}$ when one of the two sets of models is misspecified. However, because of its complex form and because in practice one does not know which model is correct, we do not derive a consistent estimator for it. We show in simulations that the variance estimator (3.15), derived under the assumption of both sets of models being correct, is somehow robust to model misspecification. Alternatively we suggest the use of bootstrap. The use of bootstrap is typical of the doubly robust literature (Zhang and Schaubel, 2012a; Bai et al., 2017; Yang et al., 2020) and the validity of such procedure is due to the fact that $\hat{\beta}$

is asymptotically linear. We will use a standard nonparametric bootstrap, where one draws bootstrap samples from $(X_i, \delta_i, A_i, Z_i)$, $i = 1, \ldots, n$ with replacement.

Detailed proofs of the Theorems are contained in Section 3.8.5 of the Supplement.

3.4.2 Asymptotic properties of Score 2.

Similarly to the previous section, we prove consistency and asymptotic normality of $\hat{\beta}^{(2)}$. The majority of the assumptions needed for studying the asymptotic behavior of $\hat{\beta}^{(2)}$ are the same needed for $\hat{\beta}^{(1)}$. Again, we need the estimators $\hat{S}_c(\cdot|\cdot,\cdot)$, $\hat{\pi}(\cdot)$ and $\hat{\Lambda}(\cdot,\cdot)$ to convergen to some $S_c^*(\cdot|\cdot,\cdot), \pi^*(\cdot), \Lambda^*(\cdot,\cdot)$, possibly different from the true quantities, as specified in Assumption 1. The following results prove that $\hat{\beta}^{(2)}$ shares the same properties of $\hat{\beta}^{(1)}$; being both rate-doubly robust and model-doubly robust.

Theorem 6. *Let Assumption 1 and Assumptions S1-S7 and S9 in the Supplement hold. If either $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$ or $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ we have $\sqrt{n}(\hat{\beta}^{(2)} - \beta_0) = o_p(1)$.*

Theorem 7. *Let Assumption 1 and Assumptions S1-S7 and S9 in the Supplement hold.*

a) (Model-double robustness): Let $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot) = \pi_0(\cdot)$ and $\Lambda^*(\cdot,\cdot) \neq \Lambda_0(\cdot,\cdot)$. Let $a_n = n^{-1/2}$ and $b_n = n^{-1/2}$; specifically let Assumption A'1 in the Supplement hold. Then $\sqrt{n}\left(\hat{\beta}^{(2)} - \beta_0\right)$ is asymptotically linear with influence function $\left\{J^{(a)}\right\}^{-1} \phi^{(a)}$ and therefore,

$$\sqrt{n}\left(\hat{\beta}^{(1)} - \beta_0\right) \xrightarrow{D} \mathcal{N}(0, \Gamma^{(a)} = \left\{E(J^{(a)})^{-1}\right\}^{\top} Var(\phi^{(a)}) E(J^{(a)})^{-1}).$$

b) (Model-double robustness): Let $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$, $S_c^*(\cdot|\cdot,\cdot) \neq S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) \neq \pi^o(\cdot)$. Let $c_n = n^{-1/2}$; specifically let Assumption B'1 in the Supplement hold. Then $\sqrt{n}\left(\hat{\beta}^{(2)} - \beta_0\right)$ is *asymptotically linear with influence function* $\left\{J^{(b)}\right\}^{-1}\phi^{(b)}$ *and therefore*

$$\sqrt{n}\left(\hat{\beta}^{(2)}-\beta_0\right)\xrightarrow{D}\mathcal{N}(0,\Gamma^{(b)}=\left\{E(J^{(b)})^{-1}\right\}^{\top}Var(\phi^{(b)})E(J^{(b)})^{-1}).$$

c) (Rate-double robustness): Let $S_c^*(\cdot|\cdot,\cdot)=S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot)=\pi_0(\cdot)$ *and* $\Lambda^*(\cdot,\cdot)=\Lambda_0(\cdot,\cdot)$ *with* $a_nc_n=o(n^{-1/2})$ *and* $b_nc_n=o(n^{-1/2})$ *and let Assumptions C'1, 2 in the Supplement hold. We have*

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{D}\mathcal{N}(0,\Gamma^{(c)}=(W^{(c')})^{-1}V^{(c')}(\tau)(W^{(c')})^{-1}),$$

where $W^{(c')}$ *is a 2x2 diagonal matrix with diagonal element* $E\left[A\int_0^X\{A-\mathcal{E}_A(t;\beta_0,S_{c0},\pi_0,Z)\}\,dt\right]$ *and* $V^{(c')}(\tau)$ *is a diagonal matrix with diagonal elements* $\int_0^\tau\left\{p'(u)\beta_{j0}+q_j'(u)\right\}du$.

Quantities $J^{(a)},J^{(b)},\phi^{(a)},\phi^{(b)}$ *are given in Section 3.8.4 of the Supplement. Quantities* p' *and* q' *are defined in Assumption C'1 in the Supplement.*

Theorem 8. *Let Assumption 1 and Assumptions S1-S7 and S9 in the Supplement hold. If* $S_c^*(\cdot|\cdot,\cdot)=S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot)=\pi_0(\cdot)$ *and* $\Lambda^*(\cdot,\cdot)=\Lambda_0(\cdot,\cdot)$ *with* $a_nc_n=o(n^{-1/2})$ *and* $b_nc_n=o(n^{-1/2})$ *and Assumptions C'1, 2 in the Supplement hold, the asymptotic variance of* $\hat{\beta}^{(2)}$ *can be consistently estimated by*

$$(\hat{W}^{(c')})^{-1}\hat{V}^{(c')}(\tau)(\hat{W}^{(c')})^{-1}, \tag{3.16}$$

where:

$$\hat{W}_{jj}^{(c')}=\frac{1}{n}\sum_{i=1}^{n}A_i\int_0^{X_i}\left\{A_i-\mathcal{E}_A(t;\hat{\beta},\hat{S}_c,\hat{\pi},Z_i)\right\}dt,$$

and

$$\hat{V}_{jj}^{(c')}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\delta_i = 1, \varepsilon_i = j\} \left\{A_i - \mathcal{E}_A(X_i; \hat{\beta}, \hat{S}_c, \hat{\pi}, Z_i)\right\}^2 .$$

The techniques used for the proof of these results are similar to the one needed for the proofs of the asymptotic results concerning $\hat{\beta}^{(1)}$. We therefore report only a sketch of the proof in Section 3.8.5 of the Supplement.

3.5 Simulation experiments

In this section we investigate the performance of our proposed estimators on a series of simulated dataset. We call $\hat{\beta}_1$ and $\hat{\beta}_2$ the results of the score (3.13) and (3.14), respectively. The first score offers a closed form solution, while for the second one, we use Newton Raphson to approximate its roots with 0 as starting point. We consider the covariates Z to be 2-dimensional. For ease of exposition we define the set $\mathcal{M} = \{h_j(t|a,z) = \beta_j a + \lambda_j(t,z) : \mathit{for\,all}\ \beta_j \in \mathbb{R}, \lambda_j(\cdot,\cdot) : [0,\tau] \times \mathbb{R}^2 \to \mathbb{R},\ j = 1,\dots,J\}$ that describes model (3.1).

3.5.1 Independent censoring

We consider four different simulation scenarios defined in Table 3.1. The censoring variable C is simulated independently of $T, A, Z = [Z_1, Z_2]^\top$; estimation of the censoring distribution is not required and we therefore use the simplified scores (3.8) and (3.9). We consider, for estimation of the propensity score and Λ_j, the following working models: $\mathcal{A}_{log} = \left\{\pi(z;\alpha) = \mathrm{expit}(\alpha^\top z) : \mathit{for\,all}\ \alpha \in \mathbb{R}^2\right\}$, $\mathcal{A}^*_{log} = \{\pi(z;\alpha) = \mathrm{expit}(\alpha^\top z + \alpha^* z_1 z_2) :$ $\mathit{for\,all}\ [\alpha, \alpha^*]^\top \in \mathbb{R}^3\}$ and $\mathcal{B} = \{\Lambda_j(t,z;G_j,\gamma_j) = G_j(t) + \gamma_j^\top z :\ \mathit{for\,all}\ \gamma_j \in \mathbb{R}^2,$

$G_j(\cdot) : [0,\tau] \to \mathbb{R}^+,\ j = 1,\dots,J\}$. Moreover, we investigate the performance of gradient boosted logistic regression (twang (Cefalu et al., 2021)) for estimation of the propensity score and we call $\mathcal{A}_{tw}$ the corresponding working model.

Both $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are consistent and asymptotically normal under model $\mathcal{M} \cap (\mathcal{A} \cup \mathcal{B})$, where $\mathcal{A}$ is the model chosen for the propensity score, $\mathcal{A}_{log}$ or $\mathcal{A}^*_{log}$ or $\mathcal{A}_{tw}$. For comparison, we also estimate β separately fitting the traditional semiparametric additive hazards model to each competing risk. We call this method traditional and we remind the reader that it is consistent under model $\mathcal{M} \cap \mathcal{B}$.

In Scenario 1 both $\mathcal{A}_{log}$ and $\mathcal{B}$ are correctly specified while in Scenario 2 $\mathcal{A}_{log}$ is misspecified since it excludes the interaction term. In Scenario 3 and 4, $\mathcal{B}$ is misspecified but $\mathcal{A}^*_{log}$ is correct. The correctness of the nonparametric model $\mathcal{A}_{tw}$ is hard to asses.

For each scenario, we simulate 500 datasets of 1000 observations. The percentage of treated subjects is $40\% - 50\%$ and the percentage of censored subjects is $10\% - 30\%$. For both estimators, model-based standard errors, (3.15) and (3.16), are used to construct 95% confidence intervals. Additionally, in Scenario 4, for the first 100 simulations, we also report the bootstrap-based standard error. To this aim we sample 100 bootstrap samples without replacement.

Results of simulations are reported in Table 3.2. Our proposed estimators exhibit consistency when either one of the two models is correctly specified. On the other hand, when model $\mathcal{B}$ is misspecified, the traditional estimator appears to be biased. The model-based standard errors are proven consistent only when both models are correct. However, in Scenario 2 and 3, they still perform well, exhibiting some level of robustness to mild departure from $\mathcal{A} \cap \mathcal{B}$. In Scenario 4, the model-based standard errors show some bias. However, bootstrap-based confidence intervals show correct coverage. The use of logistic regression and boosted logistic regression for estimation of the propensity score exhibits similar performance. We conjecture that this is due to the model-double robustness of our proposed estimators.

Table 3.1: Data-generating mechanisms of Scenarios 1-4. We define $Z = [Z_1, Z_2]^\top$. Here $C \perp (T, A, Z)$.

Scenario	Data-generating mechanism	Fitted models
1	$Z_1, Z_2 \sim U(0, 0.5)$ $\text{logit}\{\pi(Z)\} = Z_1 - Z_2$ $\lambda_j(t) = 0.1A + 1 + Z_1 + Z_2$ $C \sim U(0, 3)$	$\mathcal{A}_{log}$: CORRECT and $\mathcal{A}_{tw}$ $\mathcal{B}$: CORRECT
2	$Z_1, Z_2 \sim U(0, 0.5)$ $\text{logit}\{\pi(Z)\} = 0.25(Z_1 - Z_2) - 0.5Z_1Z_2$ $\lambda_j(t) = 0.1A + 0.3 + Z_1 + Z_2$ $C \sim U(0, 3)$	$\mathcal{A}_{log}$: WRONG and $\mathcal{A}_{tw}$ $\mathcal{B}$: CORRECT
3	$Z_1 \sim \mathcal{N}(0, 1)$ $Z_2 \sim \mathcal{N}(Z_1, 1)$ $\text{logit}\{\pi(Z)\} = 0.25(Z_1 - Z_2) + 0.5Z_1Z_2 - 1$ $\lambda_j(t) = 0.1A + 0.3 + \lvert Z_1 \rvert + \log(1 + \lvert Z_2 \rvert)$ $C \sim U(0, 3)$	$\mathcal{A}^*_{log}$: CORRECT and $\mathcal{A}_{tw}$ $\mathcal{B}$: WRONG
4	$Z_1 \sim \mathcal{N}(0, 1)$ $Z_2 \sim \mathcal{N}(Z_1, 1)$ $\text{logit}\{\pi(Z)\} = 0.25(Z_1 - Z_2) + 0.5Z_1Z_2 - 1$ $\lambda_j(t) = 0.1A + \exp(Z_1 + Z_2)$ $C \sim U(0, 3)$	$\mathcal{A}^*_{log}$: CORRECT and $\mathcal{A}_{tw}$ $\mathcal{B}$: WRONG

Table 3.2: Results of simulations from Scenarios 1-4. Here $C \perp (T, A, Z)$. Column PS indicates the working model used to estimate the propensity score. For Scenario 4, the first SE and CP are model based, while the second one uses bootstrap. SD, standard deviation; SE, standard error; CP, coverage of the 95% confidence interval.

Method			Score 1				Score 2					Traditional			
			Bias	SD	SE	CP	Bias	SD	SE	CP		Bias	SD	SE	CP
Scenario	PS	β									β				
1	Logistic	β_1	−0.012	0.156	0.146	0.93	−0.006	0.157	0.146	0.93	β_1	-0.006	0.157	0.147	0.93
		β_2	0.0006	0.144	0.147	0.95	0.006	0.146	0.146	0.95					
	Twang	β_1	−0.012	0.159	0.150	0.93	−0.006	0.161	0.149	0.93	β_2	0.006	0.146	0.147	0.95
		β_2	−0.0003	0.147	0.150	0.96	0.005	0.148	0.149	0.95					
2	Logistic	β_1	−0.010	0.108	0.120	0.97	−0.007	0.108	0.120	0.97	β_1	-0.007	0.108	0.121	0.97
		β_2	0.001	0.127	0.121	0.95	0.005	0.129	0.121	0.95					
	Twang	β_1	−0.011	0.110	0.124	0.97	−0.008	0.110	0.124	0.97	β_2	0.005	0.128	0.121	0.95
		β_2	0.001	0.131	0.124	0.95	0.004	0.133	0.124	0.94					
3	Logistic	β_1	−0.009	0.160	0.163	0.96	0.001	0.162	0.163	0.95	β_1	0.336	0.170	0.163	0.48
		β_2	0.000	0.157	0.163	0.97	0.011	0.158	0.163	0.96					
	Twang	β_1	−0.006	0.158	0.162	0.97	0.006	0.161	0.162	0.96	β_2	0.350	0.163	0.163	0.42
		β_2	0.006	0.153	0.162	0.97	0.018	0.155	0.163	0.96					
4	Logistic	β_1	0.004	0.091	0.077 0.099	0.90 0.98	0.006	0.091	0.080 0.099	0.91 0.98	β_1	0.570	0.127	0.095	0
		β_2	0.002	0.094	0.077 0.099	0.89 0.96	0.003	0.095	0.080 0.099	0.90 0.96					
	Twang	β_1	0.044	0.089	0.075 0.095	0.84 0.93	0.047	0.092	0.079 0.100	0.86 0.95	β_2	0.566	0.128	0.095	0
		β_2	0.041	0.092	0.075 0.095	0.89 0.90	0.044	0.095	0.079 0.101	0.86 0.93					

3.5.2 Dependent censoring

To investigate the robustness of our proposed estimators with respect to the censoring distribution, we consider the same settings as Scenario 1,2,3 but with censoring dependent on the covariates A, Z. Specifically we consider 4 different scenarios defined in Table 3.3.

We report the results of using both the simplified scores, (3.8) and (3.9), under the assumption of independent censoring ($\hat{\beta}^{(1)}, \hat{\beta}^{(2)}$) and score (3.13) with an estimator for S_c plugged in ($\hat{\beta}^{(1c)}$). For estimation of the censoring distribution, we consider the following working model: $\mathcal{C} = \left\{ S_c(t|a,z;\eta,\Lambda_c) = \exp\left\{ -\Lambda_c(t) e^{\eta^\top d} \right\} : \; for\, all\; \eta \in \mathbb{R}^3, \Lambda_c(\cdot) : [0,\tau] \to \mathbb{R}^+ \right\}$ where $D = [A, Z]^\top$.

$\hat{\beta}^{(1c)}$ is consistent and asymptotically normal under model $\mathcal{M} \cap \{(\mathcal{A} \cap \mathcal{C}) \cup \mathcal{B}\}$.

In Scenario 5 both $\mathcal{A}_{log} \cap \mathcal{C}$ and $\mathcal{B}$ are correctly specified while in Scenario 6, $\mathcal{A}_{log}$ is misspecified since it excludes the interaction term. In Scenario 7 both $\mathcal{A}_{log}$ and $\mathcal{C}$ are misspecified while in Scenario 8, $\mathcal{B}$ is misspecified but $\mathcal{A}^*_{log} \cap \mathcal{C}$ is correct.

For each scenario, we simulate 500 datasets of 1000 observations. The percentage of treated subjects is $40\% - 50\%$ and the percentage of censored subjects is $10\% - 30\%$. For both estimators, the model-based standard errors, (3.15) and (3.16), are used to construct 95% confidence intervals.

Results of simulations are reported in Table 3.4. In every scenario, the proposed estimators are unbiased. This seems to suggest that both our proposed scores are not really sensitive to the censoring distribution. However, Scenario 7 shows that the model-based standard error is somehow sensitive to departure from the censoring model. The user needs to carefully construct the censoring model if s/he intends to use the model-based approach.

Table 3.3: Data-generating mechanisms of Scenarios 5-8. We define $Z = [Z_1, Z_2]^\top$. Here $C \perp T | A, Z$.

Scenario	Data-generating mechanism	Fitted models
	$Z_1, Z_2 \sim U(0, 0.5)$	
5	$\text{logit}\{\pi(Z)\} = Z_1 - Z_2$	$\mathcal{A}_{log}$: CORRECT and $\mathcal{A}_{tw}$
	$C \sim Exp(\exp(-1 + A + Z_1 + Z_2))$	$\mathcal{C}$: CORRECT
	$\lambda_j(t) = 0.1A + 1 + Z_1 + Z_2$	$\mathcal{B}$: CORRECT
	$Z_1, Z_2 \sim U(0, 0.5)$	
6	$\text{logit}\{\pi(Z)\} = 0.25(Z_1 - Z_2) - 0.5 Z_1 Z_2$	$\mathcal{A}_{log}$: WRONG and $\mathcal{A}_{tw}$
	$C \sim Exp(\exp(-1 + A + Z_1 + Z_2))$	$\mathcal{C}$: CORRECT
	$\lambda_j(t) = 0.1A + 0.3 + Z_1 + Z_2$	$\mathcal{B}$: CORRECT
	$Z_1, Z_2 \sim U(0, 0.5)$	
7	$\text{logit}\{\pi(Z)\} = 0.25(Z_1 - Z_2) - 0.5 Z_1 Z_2$	$\mathcal{A}_{log}$: WRONG and $\mathcal{A}_{tw}$
	$\lambda_c(t \vert A, Z) = 2t + A - Z_1 - Z_2$	$\mathcal{C}$: WRONG
	$\lambda_j(t) = 0.1 * A + 0.3 + Z_1 + Z_2$	$\mathcal{B}$: CORRECT
	$Z_1 \sim \mathcal{N}(0, 1)$	
	$Z_2 \sim \mathcal{N}(Z_1, 1)$	
8	$\text{logit}\{\pi(Z)\} = 0.25(Z_1 - Z_2) + 0.5 Z_1 Z_2 - 1$	$\mathcal{A}^*_{log}$: CORRECT and $\mathcal{A}_{tw}$
	$C \sim Exp(\exp(-A + Z_1 - Z_2))$	$\mathcal{C}$: CORRECT
	$\lambda_j(t) = 0.1A + 0.3 + \vert Z_1 \vert + \log(1 + \vert Z_2 \vert)$	$\mathcal{B}$: WRONG

Table 3.4: Results of simulations from Scenarios 5-8. Column PS indicates the working model used to estimate the propensity score. Here $C \perp T | A, Z$. SD, standard deviation; SE, standard error; CP, coverage of the 95% confidence interval.

Method			Score 1				Score 1-Cens				Score 2			
			Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
Scenario	PS	β												
5	Logistic	β_1	−0.021	0.157	0.181	0.97	−0.032	0.163	0.180	0.97	−0.006	0.154	0.180	0.98
		β_2	−0.018	0.160	0.182	0.97	−0.026	0.168	0.181	0.96	−0.003	0.158	0.180	0.97
	Twang	β_1	−0.021	0.159	0.186	0.97	−0.030	0.166	0.184	0.97	−0.007	0.157	0.184	0.97
		β_2	−0.018	0.163	0.186	0.97	−0.026	0.171	0.185	0.95	−0.003	0.162	0.184	0.97
6	Logistic	β_1	−0.009	0.115	0.113	0.95	−0.012	0.115	0.119	0.96	−0.006	0.115	0.113	0.95
		β_2	−0.011	0.127	0.113	0.92	−0.013	0.124	0.119	0.94	−0.008	0.127	0.113	0.93
	Twang	β_1	−0.009	0.120	0.117	0.95	−0.013	0.118	0.123	0.96	−0.005	0.119	0.117	0.95
		β_2	−0.013	0.131	0.116	0.91	−0.014	0.128	0.123	0.93	−0.009	0.131	0.116	0.91
7	Logistic	β_1	−0.007	0.126	0.136	0.96	−0.012	0.128	0.336	0.83	0.000	0.127	0.135	0.95
		β_2	−0.008	0.128	0.136	0.97	−0.014	0.131	0.368	0.82	−0.002	0.129	0.135	0.97
	Twang	β_1	−0.007	0.128	0.140	0.96	−0.012	0.129	0.339	0.83	0.001	0.129	0.139	0.97
		β_2	−0.009	0.135	0.140	0.96	−0.014	0.138	0.362	0.81	−0.003	0.135	0.139	0.96
8	Logistic	β_1	0.001	0.170	0.160	0.93	−0.030	0.184	0.197	0.96	0.012	0.168	0.161	0.94
		β_2	0.002	0.164	0.160	0.95	−0.034	0.177	0.197	0.97	0.013	0.163	0.161	0.94
	Twang	β_1	0.011	0.168	0.159	0.94	−0.024	0.185	0.197	0.96	0.020	0.166	0.160	0.95
		β_2	0.014	0.160	0.159	0.95	−0.026	0.176	0.198	0.97	0.022	0.160	0.159	0.94

3.6 Application

We study the effect of mid-life alcohol consumption on the development of late-life dementia related to Alzheimer disease. To this aim we use data from the linked epidemiologic projects Honolulu Hearth Program (HHP) and Honolulu-Asia Aging Study (HAAS). HHP was established in 1965 as epidemiologic study of rates and risk factors for heart disease and stroke in men of Japanese ancestry living in Oahu and born between 1900 and 1919. 8006 men participated in the initial examination and interview (1965, then aged 45-65 years). HHP comprises of 2 further exams, (exam 2, n=7498, 1968-1971), (exam 3, n=6860, 1971-74) and a subsequent follow-up interview (mailout, n=4655, 1986-89). HAAS was established in 1991 (HHP exam 4, n=3734) as a continuation of the HHP with a shift focus on brain aging, AS, vascular dementia, other causes of cognitive and motor impairment, stroke, and the common chronic conditions of late-life. Eight further exams were done at 2-3 years intervals until 2012. During all the 9 HAAS examinations neuropsychological screenings were performed.

Here we study the effect of mid-life alcohol exposure on late-life development of moderate cognitive impairment. Cognitive impairment is assessed through the score of Cognitive Assessment and Screening Instrument (CASI), collected on the participants starting from exam 4. A score below 74 is considered moderate impairment. The mid-life alcohol exposure was assessed by self report and translated into units of drinks per month at exam 1 and exam 3.

At the end of the study, only about 500 of the 8006 men were still alive and so death without development of cognitive impairment is a competing risk for the event of interest. Since our focus is on the cognitive impairment, we consider exam 4 as time 0 and we restrict the analysis to the set of participants who had normal cognitive functions at exam 4. After deleting 30 observations with missing entries, we are left with 1881 observations.

We divide the observations into people with a light exposure to alcohol both at exam 1 and 3 and people that, at least in one of the two exams, had an heavy alcohol exposure; 1390 observations are categorized as light drinkers, while 491 as heavy drinkers. Of the 1390 in the first group, 557 developed cognitive impairment by the end of the study while 474 died, of the 491 in the second group, 216 developed cognitive impairment and 163 died. The cumulative incidence function curves for the two groups are presented in Figure 3.1.

The baseline covariates used to adjust for confounding are systolic blood pressure, age, maximum years of education, ApoE genotype and heart rate. A table with the distribution of the baseline covariates across the two groups can be found in Table 3.5.

In order to utilize our proposed scores we estimate the propensity score both using logistic and boosted logistic regression. The distribution of the estimated propensity scores for both groups is plotted in Figure 3.4.

We utilize both proposed scores to estimate the effect of mid-life alcohol exposure on the development of moderate cognitive impairment and on the competing risk death without cognitive impairment. The value of CASI at exam 4 represents a mediator for the effect under study. Following the literature on mediation analysis, the total effect of the exposure on the outcome of interest can be decomposed into direct and indirect effect (Lange and Hansen, 2011; VanderWeele, 2011). The former is the effect of alcohol exposure on development of cognitive impairment not mediated by the value of CASI at exam 4. The latter is instead the effect that can be attributed to the value of CASI at exam 4. On the other hand we conjecture that the value of CASI at exam 4 does not mediate the effect of alcohol exposure on the competing risk death. Here, we exploit our proposed scores to compute both the total and the direct effect not including and including, respectively the mediator as covariate.

In Figure 3.5 we plot Kaplan-Meier censoring survival curves for different groups, defined by the exposure and the covariates. The plots seem to suggest that the stronger assumption of $C \perp (T,A)|Z$ does not hold here. We report the estimates $\hat{\beta}^{(1)},\hat{\beta}^{(2)},\hat{\beta}^{(1c)}$ as in subsection 3.5.2. We estimate the censoring distribution according to the Cox model.

Results of the analysis are reported in Table 3.6.

The results seem to indicate that mid-life alcohol exposure has a significant effect on both the development of cognitive impairment and death without cognitive impairment. This seems to be in line with Figure 3.1. The estimated total and direct effect of the exposure on the outcome of interest are around 0.013 and 0.009, respectively. There is no a big difference between the estimates of the total and the direct effect of the exposure on the competing risk death without cognitive impairment. Both estimated effects are indeed around 0.012. This corroborates our conjecture: the value of CASI at exam 4 does not seem to mediate the effect of alcohol exposure on death without cognitive impairment.

Table 3.5: Participants' characteristic of the HHP-HAAS dataset. Presented are mean (standard deviation) for the continuous variables, and frequency (%) for the categorical variables.

	HeavyDrinkers	*LightDrinkers*
	(n = 491)	(n = 1390)
SystolicBP	151.40 (21.95)	148.76 (21.26)
Age	77.12 (3.75)	77.05 (3.80)
Education (in years)	10.42 (2.97)	11.22 (3.12)
ApoE genotype (yes)	105 (21.4%)	254 (18.3%)
HeartRate (in 30 secs)	31.88 (4.83)	31.22 (4.62)

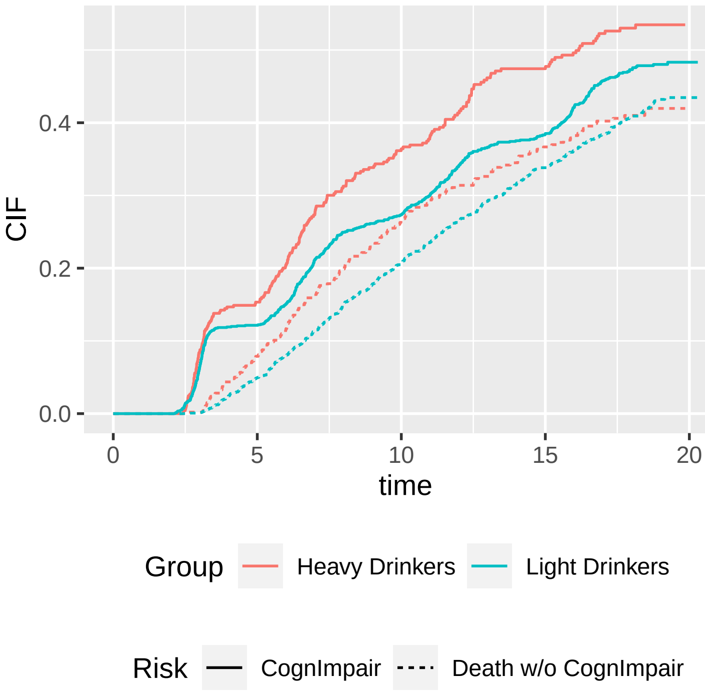


Figure 3.1: Cumulative incidence function curves for competing risks for the HHP-HAAS dataset. 'CognImpair' stands for 'Cognitive Impairment'.

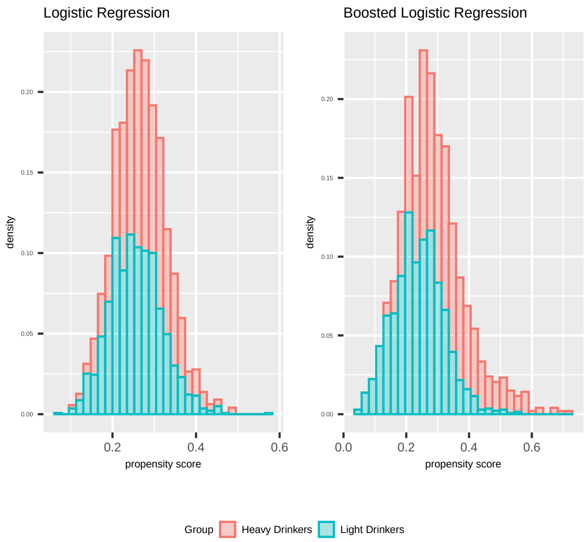


Figure 3.2: Propensity score for estimation of the total effect.

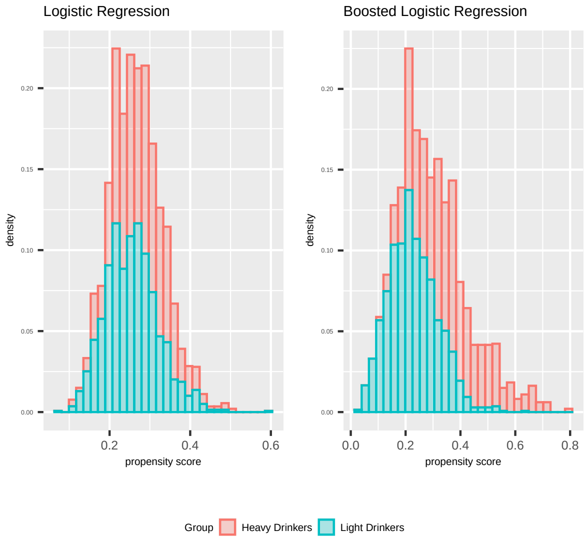


Figure 3.3: Propensity score for estimation of the direct effect.

Figure 3.4: Distribution of the estimated propensity score for the HHP-HAAS dataset.

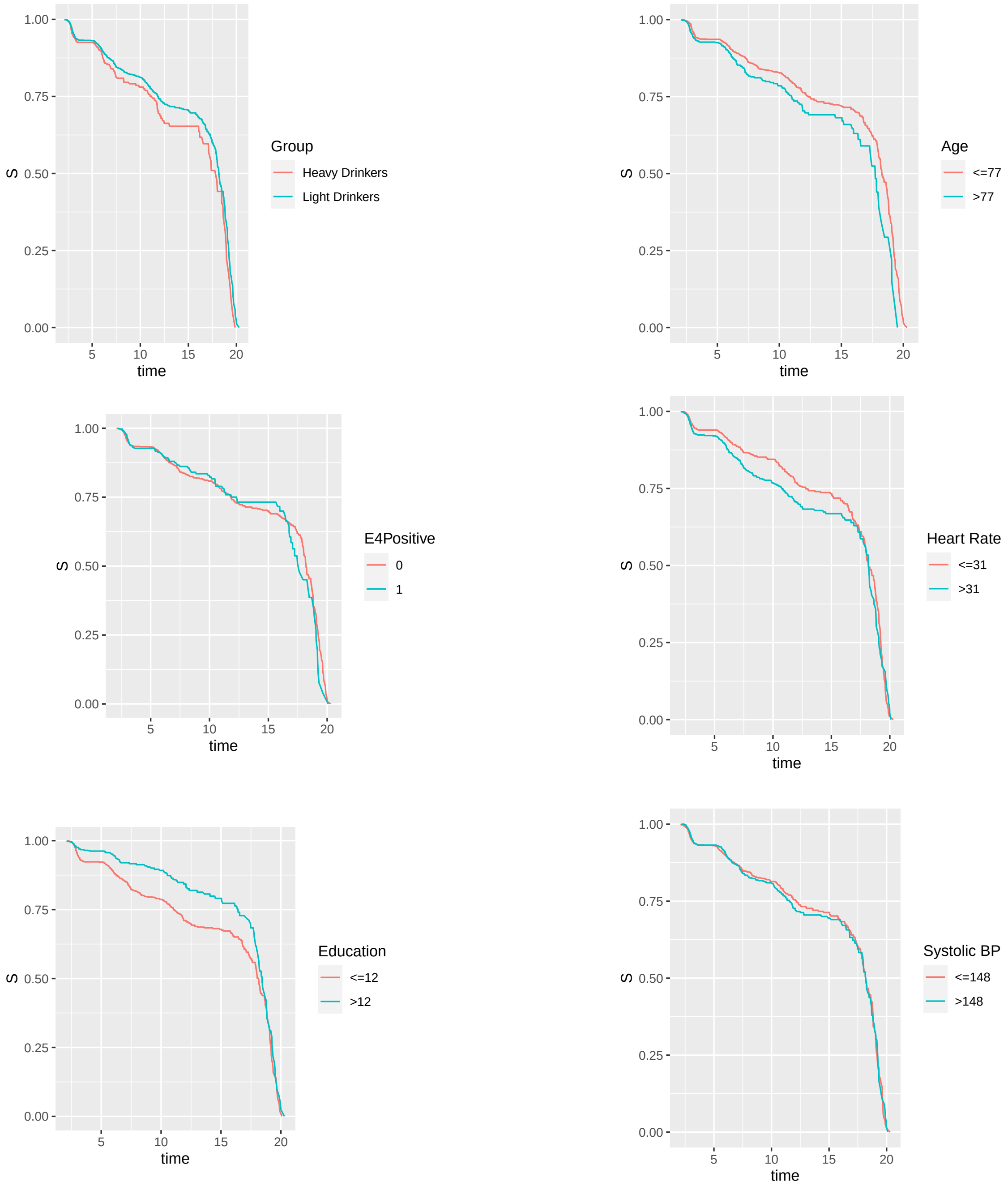


Figure 3.5: Censoring distribution for different groups of the HHP-HAAS dataset.

Table 3.6: Estimated treatment effect for the HHP-HAAS dataset. Column PS indicates the working model used to estimate the propensity score. The first CI is model-based, while the second one uses bootstrap with B=50. Time is measured in years. CI, confidence interval

Method			Score 1		Score 1-Cens		Score 2	
			$\hat{\beta}$	CI	$\hat{\beta}$	CI	$\hat{\beta}$	CI
Treatment Effect	PS	β						
Total Effect	Logistic	β_1	0.013	[0.004,0.022] [0.003,0.023]	0.015	[0.005,0.025] [0.004,0.022]	0.012	[0.004,0.021] [0.003,0.022]
		β_2	0.012	[0.005,0.020] [0.005,0.020]	0.013	[0.004,0.022] [0.005,0.021]	0.012	[0.005,0.019] [0.006,0.018]
	Twang	β_1	0.012	[0.003,0.020] [0.001,0.022]	0.013	[0.004,0.023] [0.002,0.025]	0.011	[0.003,0.020] [0.002,0.021]
		β_2	0.012	[0.004,0.020] [0.002,0.021]	0.012	[0.003,0.022] [0.001,0.024]	0.011	[0.004,0.019] [0.003,0.020]
Direct Effect	Logistic	β_1	0.010	[0.001,0.019] [0.002,0.019]	0.009	[0.000,0.019] [0.000,0.019]	0.009	[0.001,0.018] [0.002,0.019]
		β_2	0.012	[0.005,0.020] [0.005,0.022]	0.013	[0.004,0.022] [0.004,0.023]	0.012	[0.005,0.019] [0.004,0.021]
	Twang	β_1	0.008	[0.000,0.017] [0.005,0.011]	0.008	[-0.002,0.017] [-0.001,0.006]	0.008	[0.000,0.016] [0.001,0.016]
		β_2	0.010	[0.003,0.018] [0.007,0.014]	0.011	[0.002,0.020] [0.000,0.007]	0.010	[0.003,0.018] [0.003,0.018]

3.7 Discussion

The proposed estimators are model-doubly robust; they are consistent and asymptotically normal if either one of the two sets of models is correctly specified. The estimators are also rate-doubly robust, i.e. when both sets of models are correct, they only need the product of their rates of convergence to be $o(\sqrt{n})$. This characteristic allows the user to choose, for estimation of the nuisance parameters, from a big variety of methodologies, both parametric and nonparametric. To the best of our knowledge we provide the first doubly robust estimators for the hazard difference in the presence of competing risks.

In this article we have proposed two doubly robust estimators for the conditional cause-specific hazard difference under competing risks. We proposed two estimators that are model-doubly robust: they are consistent and asymptotically normal if either both the propensity score and the censoring distribution or the outcome models for competing risks are modeled correctly. Moreover, they are rate-doubly robust: they are consistent and asymptotically normal if both sets of models are correctly specified and the product of their convergence rates is $o(\sqrt{n})$. The last property, that has its roots in the orthogonality of the scores, gives the user the possibility to estimate the propensity score or/and the censoring distribution or/and the outcome model using modern nonparametric methods, known to have rates of convergence slower than $\sqrt{n}$. In simulations we showed the performance of our estimators when boosted logistic regression is used for estimation of the propensity score. Different nonparametric methods are also available for estimation of survival curves, such as survival random forest (Ishwaran et al., 2008), spline (Gray, 1992; Kooperberg et al., 1995a) and Kernel (Beran, 1981; Dabrowska, 1989). The user can employ them for estimation of $\Lambda_j(t,Z)$ and then use one of our proposed scores. In the absence of competing risks, Hou et al. (2021) proposed in section 6.1 to estimate nonparametrically the cumulative hazard function separately for the treated and the untreated, $\hat{\Lambda}^{(1)}, \hat{\Lambda}^{(0)}$ and then use $\hat{\Lambda}(t,Z) = w(t)\left\{\hat{\Lambda}^{(1)}(t|Z) - \beta t\right\} + \{1 - w(t)\}\hat{\Lambda}^{(0)}(t|Z)$ for some weight $w(t)$. In our case this method needs to be adapted to the setting of competing risks. To this aim the user needs to carefully make use of nonparametric methods that work under such setting to estimate $\hat{\Lambda}_j^{(1)}, \hat{\Lambda}_j^{(0)}$. For example Ishwaran et al. (2014) proposed survival random forest for competing risks for estimation of both cumulative cause-specific hazard functions and cumulative incidence functions.

Here we propose estimators for the total treatment effect on both the event of interest and the competing event. Since in our example mid-life alcohol consumption has an harmful effect on both the development of cognitive impairment and survival, the total effect has not a difficult interpretation. However, when the total effects on two competing risks have opposite directions, it might be difficult to understand which part of the effect is due to the competing event. If for example we had discovered a beneficial effect of heavy alcohol consumption on survival, we would have not known if the harmful effect of drinking on the development of cognitive impairment were simply a consequence of aging. To shed light on this, recently separable direct and indirect effects have been introduced (Stensrud et al., 2020). While they offer better overall interpretation, they need to be justified by subject-matter knowledge and their identification is based on untestable assumptions. The decomposition into separable effects of our total effect is beyond the scope of this paper and we leave this for future work.

In simulations we have investigated the use of both logistic regression and boosted logistic regression for estimation of the propensity score. The latter has been recently acquired popularity among practitioners. The idea that nonparametric methods are always consistent is a common misconception. While it is true that these methods relax the modeling assumptions typical of parametric methodologies, their consistency is not granted and it is often hard to asses. Moreover, nonparametric methods have often convergence rate slower than the classical $\sqrt{n}$ and their tuning process can be non trivial. Because of all the above reasons it is useful to use them in combination with estimators that are both model and rate-doubly robust as our proposals.

To use our scores, the censoring distribution needs to be modeled. However this is not true if one is willing to assume that the censoring is independent of the treatment and the failure time given the covariates. In simulations we have shown that our estimators seem to be pretty robust with respect to this assumption. However, we advise the user to asses the validity of the assumption as we have done in the data analysis. Moreover, when the censoring distribution is estimated and the model-based confidence interval is considered, the censoring model needs to be carefully constructed. To this aim, the user should choose the model that seems to best fit the data. Alternatively, if computational time is not a concern, one can use bootstrap.

3.8 Appendix

3.8.1 Derivation of the semiparametrically efficient score.

We have, for an individual, the following likelihood:

$$\begin{aligned}
L = &\prod_{j=1}^{J} \left\{h_j(X|A,Z)\right\}^{\mathbb{1}\{\delta=1,\varepsilon=j\}} \exp\left\{-H_j(X|A,Z)\right\} \left\{\lambda_c(X|A,Z)\right\}^{1-\delta} \\
&\times \exp\left\{-\Lambda_c(X|A,Z)\right\} P(A|Z) f(Z) \\
= &\prod_{j=1}^{J} \left\{\lambda_j(X,Z)+\beta_j A\right\}^{\mathbb{1}\{\delta=1,\varepsilon=j\}} \exp\left\{-\Lambda_j(X,Z)-\beta_j AX\right\} \left\{\lambda_c(X|A,Z)\right\}^{1-\delta} \\
&\times \exp\left\{-\Lambda_c(X|A,Z)\right\} P(A|Z) f(Z), \qquad (3.17)
\end{aligned}$$

where $\lambda_c(t|a,z)$ is the conditional hazard function for C, $P(a|z)$ is the conditional distribution of A and $f(z)$ is the density of the covariates. Moreover $\Lambda_j(t,z) = \int_0^t \lambda_j(u,z)du$ for $j = 1,\ldots,J$ and $\Lambda_c(t|a,z) = \int_0^t \lambda_c(u|a,z)du$.

3.8.2 Score for β

We first derive the score for the parameter of interest β. We first prove a generic result.

Lemma 3. *For a generic cause-specific hazards model $h_j(t|W;\theta)$, $j = 1,\ldots,J$ where $\theta = (\beta,\eta)$ and the parameter of interest $\beta = [\beta_1,\ldots,\beta_J]^\top$ is finite-dimensional and W are covariates, we have:*

$$S_\beta = \left\{ \int_0^\tau \partial_{\beta_j} h_j(t|W;\theta)\Big|_{\beta=\beta_0} \frac{dM_j(t)}{h_j(t|W;\theta)} \right\}_{j=1}^{J}. \qquad (3.18)$$

Proof of Lemma 3. Under a generic cause-specific hazards model the log likelihood for an individual is

$$\begin{aligned}\log L(\theta) &= \sum_{j=1}^{J}\left[\mathbb{1}\{\delta=1,\varepsilon=j\}\log\left\{h_j(X|W;\theta)\right\}-H_j(X|W;\theta)\right]\\ &\quad+(1-\delta)\log\{\lambda_c(X|A,Z)\}-\Lambda_c(X|A,Z)+\log P(A|Z)+\log f(Z),\end{aligned}$$

and the associated martingales are:

$$M_j(t)=N_j(t)-H_j(t|W;\theta_0)Y(t). \tag{3.19}$$

Therefore, for $j=1,\ldots,J$:

$$\begin{aligned}\left\{S_\beta\right\}_j &= \left.\frac{\partial\log L(\theta)}{\beta_j}\right|_{\theta=\theta_0}\\ &= \sum_{j=1}^{J}\mathbb{1}\{\delta=1,\varepsilon=j\}\left[\left.\frac{\partial_{\beta_j}h_j(X|W;\theta)}{h_j(X|W;\theta)}\right|_{\theta=\theta_0}-\left.\partial_{\beta_j}H_j(X|W;\theta)\right|_{\theta=\theta_0}\right]\\ &= \int_0^\tau\left.\frac{\partial_{\beta_j}h_j(t|W;\theta)}{h_j(t|W;\theta)}\right|_{\theta=\theta_0}dN_j(t)-\int_0^\tau\left.\partial_{\beta_j}h_j(t|W;\theta)\right|_{\theta=\theta_0}Y(t)dt\\ &= \int_0^\tau\left.\partial_{\beta_j}h_j(t|W;\theta)\right|_{\theta=\theta_0}\frac{dM_j(t)}{h_j(t|W;\theta)}.\end{aligned}$$

□

Application of the above Lemma to model (3.1) leads to:

$$S_\beta=\left\{\int_0^\tau A\frac{dM_j(t)}{h_j(t|A,Z)}\right\}_{j=1}^{J}. \tag{3.20}$$

Nuisance tangent space

Under model (3.1) we have $J+2$ nuisance parameters: $\lambda_1(t,z),\dots,\lambda_J(t,z),\lambda_c(t|a,z)$, $P(a|z)f(z)$. We call their tangent spaces $\Lambda^1_{1s},\dots,\Lambda^J_{1s},\Lambda_{2s},\Lambda_{3s}$, respectively.

Lemma 5.1 of Tsiatis (2007) proves that:

$$\Lambda_{2s}=\left\{\int_0^\tau g(t,A,Z)dM_c(t) \quad : \; for\,all\; g(t,A,Z)\right\}, \tag{3.21}$$

where $M_c(t)$ is the martingale associated with the censoring distribution. Pag. 117 of Tsiatis (2007) proves that:

$$\Lambda_{3s}=\{g(A,Z) \;:\; \mathrm{E}\{g(A,Z)\}=0\}. \tag{3.22}$$

We now derive Λ^j_{1s} for $j=1,\dots,J$.

Lemma 4. *For $j=1,\dots,J$:*

$$\Lambda^j_{1s}=\left\{\int_0^\tau g(t,Z)\frac{dM_j(t)}{h_j(t|A,Z)} \;:\; for\,all\; g(t,Z)\right\}. \tag{3.23}$$

Moreover $\Lambda^l_{1s}\perp\Lambda^j_{1s}$ for each $l\neq j$.

Proof of Lemma 4. The nuisance tangent space, when the nuisance parameter has finite dimension, is defined as the space spanned by the nuisance score. The nuisance tangent space for a semiparametric model is the mean-square closure of all parametric submodel nuisance tangent spaces. We therefore starts considering parametric submodels. Let's assume that j is fixed and let's consider a generic parametric submodel:

$$h_j(t|A,Z;\eta) = \lambda_j(t,Z;\eta) + \beta_j A,$$

where η_0 indicates the true parameter. For this parametric submodel, by Lemma 3, we have

$$S_\eta = \int_0^\tau \partial_\eta \lambda_j(t,Z;\eta)|_{\eta=\eta_0} \frac{dM_j(t)}{h_j(t|A,Z)}.$$

We hence conjecture that, for our semiparametric model:

$$\Lambda_{1s}^j = \left\{ \int_0^\tau g(t,Z) \frac{dM_j(t)}{h_j(t|A,Z)} \;\; : \;\; for\, all\; g(t,Z) \right\}. \tag{3.24}$$

By above calculations, we know that, the nuisance tangent space of any parametric submodel belongs to Λ_{1s}^j. To complete our proof we need to prove that for any element of the conjectured (10), indexed by $g(t,Z)$, there exists a parametric submodel such that, such element belongs to its nuisance tangent space. Given $g(t,Z)$, straightforward algebra proves that the score of the following parametric submodel:

$$h_j(t|A,Z;\eta) = \lambda_j(t,Z;\eta_0) + \eta g(t,Z) + \beta_j A,$$

corresponds to the element of Λ_{1s}^j indexed by the chosen $g(t,Z)$. Our conjecture is therefore proven right.

We now focus on proving the orthogonality of these spaces. For each $g_l(t,Z), g_j(t,Z)$ with

$l \neq j$, we have:

$$\begin{aligned} & \mathrm{E}\left\{\int_0^\tau g_l(t,Z)\frac{dM_l(t)}{h_l(t \mid A,Z)} \times \int_0^\tau g_j(t,Z)\frac{dM_j(t)}{h_j(t \mid A,Z)}\right\} \\ = \; & \mathrm{E}\left\{\int_0^\tau g_l(t,Z)g_j(t,Z)\frac{1}{h_l(t \mid A,Z)h_j(t \mid A,Z)} < dM_l(t), dM_j(t) > \right\} = 0, \end{aligned}$$

where the last equality comes from the fact that we assume that competing risks don't happen at the same time. Therefore $\Lambda_{1s}^l \perp \Lambda_{1s}^j$ for $l \neq j$. □

The nuisance tangent space is therefore:

$$\Lambda = \Lambda_{1s}^1 \oplus \ldots \oplus \Lambda_{1s}^J \oplus \Lambda_{2s} \oplus \Lambda_{3s}. \tag{3.25}$$

Orthogonal complement of the nuisance tangent space: proof of Lemma 1

We are now ready to derive the orthogonal complement of the nuisance tangent space. We start with some useful lemma:

Lemma 5. *For any $g_j(t,A,Z)$:*

$$\prod\left\{\int_0^\tau g_j(t,A,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\middle|\Lambda_{1s}^j\right\} = \int_0^\tau g_j^*(t,Z)\frac{dM_j(t)}{h_j(t|A,Z)},$$

where

$$g_j^*(t,Z) \;\; = \;\; \frac{E\left[g_j(t,A,Z)h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^J \beta_l A t}\middle|Z\right]}{E\left[h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^J \beta_l A t}\middle|Z\right]}.$$

Proof of Lemma 20. By definition of projection, we need, for any $g(t,Z)$, that:

$$\begin{aligned}
0 &= \mathrm{E}\left[\int_0^\tau \left\{g_j(t,A,Z)-g_j^*(t,Z)\right\}\frac{dM_j(t)}{h_j(t|A,Z)} \times \int_0^\tau g(t,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\right] \\
&= \mathrm{E}\left[\int_0^\tau \left\{g_j(t,A,Z)-g_j^*(t,Z)\right\} g(t,Z)h_j^{-2}(t|A,Z) < dM_j(t) >\right] \\
&= \mathrm{E}\left[\int_0^\tau \left\{g_j(t,A,Z)-g_j^*(t,Z)\right\} g(t,Z)h_j^{-1}(t|A,Z)Y(t)dt\right] \\
&= \int_0^\tau \mathrm{E}\Big(\mathrm{E}\left[\left\{g_j(t,A,Z)-g_j^*(t,Z)\right\} h_j^{-1}(t|A,Z)Y(t)|Z\right] g(t,Z)\Big)\,dt,
\end{aligned}$$

implying that, almost surely,

$$\mathrm{E}\left[\left\{g_j(t,A,Z)-g_j^*(t,Z)\right\} h_j^{-1}(t|A,Z)Y(t)|Z\right] = 0.$$

By contradiction, let's assume that the above expectation is not zero on an interval with positive measure. If we take

$$g(t,Z) = \mathrm{E}\left[\left\{g_j(t,A,Z)-g_j^*(t,Z)\right\} h_j^{-1}(t|A,Z)Y(t)|Z\right],$$

then

$$\int_0^\tau \mathrm{E}\Big(\mathrm{E}\left[\left\{g_j(t,A,Z)-g_j^*(t,Z)\right\} h_j^{-1}(t|A,Z)Y(t)|Z\right] g(t,Z)\Big)\,dt \neq 0.$$

and so the contradiction.

Therefore:

$$\begin{aligned} g_j^*(t,Z) &= \frac{\mathrm{E}\left\{g_j(t,A,Z)h_j^{-1}(t|A,Z)Y(t)|Z\right\}}{\mathrm{E}\left\{h_j^{-1}(t|A,Z)Y(t)|Z\right\}} \\ &= \frac{\mathrm{E}\left[g_j(t,A,Z)h_j^{-1}(t|A,Z)E\left\{Y(t)|A,Z\right\}|Z\right]}{\mathrm{E}\left[h_j^{-1}(t|A,Z)E\left\{Y(t)|A,Z\right\}|Z\right]} \\ &= \frac{\mathrm{E}\left\{g_j(t,A,Z)h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}{\mathrm{E}\left\{h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\sum_{l=1}^{J}\beta_l A t}|Z\right\}}. \end{aligned}$$

□

Lemma 6. *We have:*

$$\begin{aligned} \Lambda^{\perp} &= \left\{\sum_{j=1}^{J}\int_0^{\tau}\left\{g_j(t,A,Z)-g_j^*(t,Z)\right\}\frac{dM_j(t)}{h_j(t|A,Z)}\ :\ for\,all\ g_j(t,A,Z)\right. \qquad (3.26) \\ & \left. and\ \ g_j^*(t,Z)\ \ s.t.\ \prod\left\{\int_0^{\tau} g_j(t,A,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\,\Big|\Lambda_{1s}^{j}\right\} = \int_0^{\tau} g_j^*(t,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\right\}. \end{aligned}$$

Proof of Lemma 6. The tangent space for a parametric model identified by the parameter $\theta = (\beta,\eta)$ is defined as the space spanned by the score S_θ. The tangent space for a semiparametric model is the mean-square closure of all parametric submodel tangent spaces. If we don't put any restrictions on the density that generates the data, then it follows from Theorem 4.4 of Tsiatis (2007) that the corresponding tangent space is the entire Hilbert space $\mathcal{H} = \{g(X,\delta,A,Z)\ :\ \mathrm{E}\{g\} = 0,\ \mathrm{E}\{g^{\top}g\} < \infty\}$. Model (3.1) imposes restrictions on the cause-specific hazards $h_j(t|A,Z)$ for $j = 1,\ldots,J$ and it leaves $\lambda_c(t|A,Z)$, $P(A|Z)$ and $f(Z)$ unspecified. Suppose that now we don't put any restriction and we consider a nonparametric model in which also $h_j(t|A,Z)$ are left unspecified. The tangent space for this nonparametric model, i.e. $\mathcal{H}$, is:

$$\mathcal{H} = \Lambda_{1s}^{1*} \oplus \ldots, \oplus \Lambda_{1s}^{J*} \oplus \Lambda_{2s} \oplus \Lambda_{3s}, \tag{3.27}$$

where Λ_{1s}^{j*} are the spaces associated with $h_j(t|A,Z)$; now left arbitrary. Similarly to what we have done in the proof of Lemma 4, it is easy to show that:

$$\Lambda_{1s}^{j*} = \left\{ \int_0^\tau g_j(t,A,Z) \frac{dM_j(t)}{h_j(t|A,Z)} \;:\; for\, all\; g_j(t,A,Z) \right\}, \tag{3.28}$$

and that, for any $l \neq j$:

$$\Lambda_{1s}^{l*} \perp \Lambda_{1s}^{j*},\; \Lambda_{1s}^{l*} \perp \Lambda_{1s}^{j},\; \Lambda_{1s}^{l} \perp \Lambda_{1s}^{j*}. \tag{3.29}$$

By definition, the orthogonal complement of the nuisance tangent space is

$$\Lambda^{\perp} = \{g - \Pi(g|\Lambda) \;:\; for\, each\; g \in \mathcal{H}\}. \tag{3.30}$$

We remind the reader that

$$\Lambda = \Lambda_{1s}^{1} \oplus \ldots \oplus \Lambda_{1s}^{J} \oplus \Lambda_{2s} \oplus \Lambda_{3s}. \tag{3.31}$$

By (3.27), (3.30) and (3.31), to find $\Lambda^{\perp}$ it is sufficient to find the residual of the projection of an arbitrary element of $\Lambda_{1s}^{1*} \oplus \ldots, \oplus \Lambda_{1s}^{J*}$ onto $\Lambda_{1s}^{1} \oplus \ldots, \oplus \Lambda_{1s}^{J}$.

By (3.28) and (3.29) we have, for any $g_1(t,A,Z),\ldots,g_J(t,A,Z)$:

$$\begin{aligned}
&\sum_{j=1}^{J}\prod\left\{\int_0^\tau g_j(t,A,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\right|\Lambda_{1s}^1\oplus\ldots,\oplus,\Lambda_{1s}^J\Bigg\}\\
=\ &\sum_{l=1}^{J}\sum_{j=1}^{J}\prod\left\{\int_0^\tau g_j(t,A,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\right|\Lambda_{1s}^l\Bigg\}\\
=\ &\sum_{j=1}^{J}\prod\left\{\int_0^\tau g_j(t,A,Z)\frac{dM_j(t)}{h_j(t|A,Z)}\right|\Lambda_{1s}^j\Bigg\}.
\end{aligned}$$

Expression (3.26) is therefore proven. □

By lemma 20 and 6, we can therefore conclude that:

$$\begin{aligned}
\Lambda^\perp \ = \ &\left\{\sum_{j=1}^{J}\int_0^\tau\left[g_j(t,A,Z)-\frac{\mathrm{E}\left\{g_j(t,A,Z)h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\Sigma_{l=1}^J\beta_l At}|Z\right\}}{\mathrm{E}\left\{h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\Sigma_{l=1}^J\beta_l At}|Z\right\}}\right]\right.\\
&\left.\times\frac{dM_j(t)}{h_j(t|A,Z)}:\ \ for\ each\ g_j(t,A,Z)\right\}. \qquad (3.32)
\end{aligned}$$

Efficient score: proof of Lemma 2

The efficient score is given by the projection of S_β onto $\Lambda^\perp$. By the form of S_β, its projection on $\Lambda^\perp$ would be the element of $\Lambda^\perp$ that corresponds to $g_1(t,A,Z)=Ae_1,\ldots,g_J(t,A,Z)=Ae_J$, where e_j is a 0 vector with 1 at the j^{th} position. Therefore:

$$S_{eff} \ = \ \left\{\int_0^\tau\left[A-\frac{\mathrm{E}\left\{Ah_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\Sigma_{l=1}^J\beta_l At}|Z\right\}}{\mathrm{E}\left\{h_j^{-1}(t|A,Z)S_c(t|A,Z)e^{-\Sigma_{l=1}^J\beta_l At}|Z\right\}}\right]\frac{dM_j(t)}{h_j(t|A,Z)}\right\}_{j=1}^{J}. \qquad (3.33)$$

3.8.3 Technical Assumptions

General Assumptions for Theorems 3-8

Assumption S 1. β_0 *is contained in the interior of a compact set.*

Assumption S 2. *We have a finite upper bound of time* τ *and, for* $j = 1, \ldots, J$*, there exist finite* $L_j < \infty$ *such that* $\sup_{z \in \mathcal{Z}} \Lambda_j(\tau, z) \leq L_j$.

Assumption S 3. *There are no ties, both across observations and both across events.*

Assumption S 4. *There exists* C_1*, such that* $P\left(\sup_{i=1,\ldots,n} \|Z\|_\infty \leq C_1\right) = 1$.

Assumption S 5. *There exist a positive* C_2 *such that*

$$\inf_{z \in \mathcal{Z}, a=0,1} S_{c0}(\tau|a, z) > C_2 > 0,$$

and C_3*,* C_4 *such that:*

$$0 < C_3 < \inf_{z \in \mathcal{Z}} \pi_0(z) < \sup_{z \in \mathcal{Z}} \pi_0(z) < C_4 < 1.$$

There exists a strictly positive constant $\varepsilon > 0$ *such that*

$$Var(A|Z) > \varepsilon,$$

$$E\{N(\tau)|A = 0, Z\} < 1 - \varepsilon,$$

$$E\{Y(\tau)|A = 0, Z\} > \varepsilon,$$

$$E\{Y(\tau)|A,Z\} > \varepsilon.$$

Assumption S 6. *If the estimator $\hat{\Lambda}$ depends on the unknown β for $j = 1,2$ in a neighborhood of β_0,*

$$\bigvee_{t=0}^{\tau} \sup_{z\in\mathcal{Z}} \{\hat{\Lambda}_j(t,z;\beta_j) - \hat{\Lambda}_j(t,z;\beta_{j0})\} = O_p(|\beta_j - \beta_{j0}|),$$

where $\bigvee_{t=0}^{\tau} g(t) = \sup_{0<t_0<\ldots<t_N=\tau, N\in\mathbb{N}} \sum_{j=1}^{N} \left|g(t_{j-1}) - g(t_j)\right|.$

Assumption S 7. *There exist a positive C_5 such that*

$$\inf_{z\in\mathcal{Z}, a=0,1} S_c^*(\tau|a,z) > C_5 > 0,$$

and C_6, C_7 such that:

$$0 < C_6 < \inf_{z\in\mathcal{Z}} \pi^*(z) < \sup_{z\in\mathcal{Z}} \pi^*(z) < C_7.$$

Assumption S 8. *There exists $\varepsilon > 0$ such that:*

$$E\left|\int_0^{\tau} [1 + td\{\Lambda_1^*(t,Z) - \Lambda_{10}(t,Z)\} + td\{\Lambda_2^*(t,Z) - \Lambda_{20}(t,Z)\}]\, dt\right| > \varepsilon.$$

Moreover, If the estimator $\hat{\Lambda}$ depends on the unknown β,

$$E\left[\{A - \pi^*(Z)\}\{A - E(q_j(t))\}|Z\right] > \varepsilon,$$

where we call $q_{ji}(t)$ a function, such that:

$$\hat{\Lambda}_j(t,Z;\beta) - \hat{\Lambda}_j(t,Z;\beta_0) \;\; = \;\; (\beta_j - \beta_{j0}) \times \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau q_{ji}(t).$$

Assumption S 9. *There exists* $\varepsilon > 0$ *such that*

$$\int_0^\tau E\left[A - te^{-\sum_{j=1}^{J}\beta_{j0}At} S_c(t|A,Z) d\left\{\Lambda_1^*(t,Z) - \Lambda_{10}(t,Z) + \Lambda_2^*(t,Z) - \Lambda_{20}(t,Z)\right\}|Z\right] > \varepsilon,$$

$$\int_0^\tau E\left[\left\{A + d\partial_{\beta_j}\Lambda_j^*(t,Z)\right\}\left\{A - \mathcal{E}_A(t;\beta,S_c,\pi,Z)\right\}|Z\right] > \varepsilon.$$

Assumptions for Theorem 4

In Theorem 4 we consider three different cases, depending on which model is correctly specified: Each case needs specific assumptions.

Assumption A 1. *Let, for* $j = 1,2$

$$P_{1j}^{(a)}(t) := \frac{1}{n}\sum_{i=1}^{n} e^{(\beta_{10}+\beta_{20})A_i t}\left\{A_i - \pi(Z_i;\alpha_0)\right\}\partial_\eta S_c^{-1}(t|A,Z;\eta_0,\Lambda_{c0})dM_{ji}(t;\beta_{j0},\Lambda_j^*),$$

$$P_{2j}^{(a)}(t) := \frac{1}{n}\sum_{i=1}^{n} e^{(\beta_{10}+\beta_{20})A_i t}\left\{A_i - \pi(Z_i;\alpha_0)\right\}\partial_{\Lambda_c} S_c^{-1}(t|A,Z;\eta_0,\Lambda_{c0})dM_{ji}(t;\beta_{j0},\Lambda_j^*),$$

and

$$P_{3j}^{(a)}(t) := \frac{1}{n}\sum_{i=1}^{n} e^{(\beta_{10}+\beta_{20})A_i t}\partial_\alpha \pi(Z_i;\alpha_0) S_c^{-1}(t|A,Z;\eta_0,\Lambda_{c0})dM_{ji}(t;\beta_{j0},\Lambda_j^*).$$

For $l = 1,2,3$, there exist some bounded $p_{lj}^{(a)}(t)$ and a neighborhood $\mathcal{B}$ of $\{\beta_0, S_{c0}(\cdot|\cdot,\cdot), \pi_0(\cdot), \Lambda_j^*(\cdot,\cdot)\}$, such that:

$$\sup_{t\in[0,\tau],\{\beta,S_c,\pi,\Lambda_j\}\in\mathcal{B}} \left| P_{lj}^{(a)}(t) - p_{lj}^{(a)}(t) \right| \xrightarrow{p} 0.$$

Assumption A 2. *There exist influence functions $\sigma_1, \sigma_2(\cdot), \sigma_3$ such that, for any $t \in [0,\tau]$:*

$$\hat{\eta} - \eta_0 = \frac{1}{n}\sum_{i=1}^{n} \sigma_{1i},$$

$$\hat{\Lambda}_c(t) - \Lambda_{c0}(t) = \frac{1}{n}\sum_{i=1}^{n} \sigma_{2i}(t),$$

$$\hat{\alpha} - \alpha_0 = \frac{1}{n}\sum_{i=1}^{n} \sigma_{3i}.$$

Assumption B 1. *Let,*

$$P_{1j}^{(b)}(t) := \frac{1}{n}\sum_{i=1}^{n} e^{(\beta_{10}+\beta_{20})A_i t} \left\{S_c^*(t|A_i,Z_i)\right\}^{-1} \left\{A_i - \pi^*(Z_i)\right\} Y_i(t) \partial_{\gamma_j} dL_j(t,Z;G_{j0},\gamma_{j0}),$$

$$P_{2j}^{(b)}(t) := \frac{1}{n}\sum_{i=1}^{n} e^{(\beta_{10}+\beta_{20})A_i t} \left\{S_c^*(t|A_i,Z_i)\right\}^{-1} \left\{A_i - \pi^*(Z_i)\right\} Y_i(t) \partial_{G_j} L_j(t,Z;G_{j0},\gamma_{j0}).$$

We assume that, there exist $p_{lj}^{(b)}(t)$, for $l = 1,2$ and a neighborhood $\mathcal{B}$ of

$\{\beta_0, \Lambda_{j0}(\cdot,\cdot), S_c^*(\cdot|\cdot,\cdot), \pi^*(\cdot,\cdot)\}$ *such that :*

$$\sup_{t\in[0,\tau],\{\beta,\Lambda_j,S_c,\pi\}\in\mathcal{B}} \left|P_{lj}^{(b)}(t) - p_{lj}^{(b)}(t)\right| \xrightarrow{p} 0.$$

Assumption B 2. *There exists influence functions* $\sigma_4, \sigma_5(\cdot)$ *such that, for any* $t \in [0,\tau]$, $j = 1,\dots,J$*:*

$$\hat{\gamma}_j - \gamma_{j0} = \frac{1}{n}\sum_{i=1}^{n} \sigma_{4i},$$

$$\hat{G}_j(t) - G_{j0}(t) = \frac{1}{n}\sum_{i=1}^{n} \sigma_{5i}(t).$$

Assumption C 1. *Let*

$$h(t;A,Z) = e^{(\beta_{10}+\beta_{20})At}\{S_{c0}(t|A,Z)\}^{-1}\{A - \pi_0(Z)\},$$

$$P(t) = \frac{1}{n}\sum_{i=1}^{n} h^2(t;A_i,Z_i)A_iY_i(t),$$

and

$$Q_j(t) = \frac{1}{n}\sum_{i=1}^{n} h^2(t;A_i,Z_i)\Lambda_{j0}(t,Z_i)Y_i(t).$$

We assume that, there exists $p(t), q_j(t)$ *and a neighborhood* $\mathcal{B}$ *of the true* $\{\beta_0, S_{c0}(\cdot|\cdot,\cdot), \pi_0(\cdot), \Lambda_0(\cdot,\cdot)\}$ *such that:*

$$\sup_{t\in[0,\tau],\{\beta,S_c,\pi,\Lambda\}\in\mathcal{B}} |P(t) - p(t)| \xrightarrow{p} 0.$$

and

$$\sup_{t\in[0,\tau],\{\beta,S_c,\pi,\Lambda\}\in\mathcal{B}} \left|Q_j(t)-q_j(t)\right| \xrightarrow{p} 0.$$

Assumption C 2. *Let $p(t), q_j(t)$ as in Assumption C1, then for $j=1,2$, we assume that:*

$$\int_0^\tau \left\{p(u)\beta_j + q_j(u)\right\} du > 0.$$

Assumptions for Corollary 1

Assumption A* 1. *Let, for $j=1,2$*

$$P_{1j}^{(a')}(t) := \frac{1}{n}\sum_{i=1}^n e^{(\beta_{10}+\beta_{20})A_i t}\left\{A_i - expit(\alpha_0^\top Z_i)\right\}\exp\left(\Lambda_{c0}(t)e^{\eta_0^\top D_i}\right)\Lambda_{c0}(t)e^{\eta_0^\top D_i}D_i dM_{ji}(t;\beta_{j0},\Lambda_j^*),$$

$$P_{2j}^{(a')}(t) := \frac{1}{n}\sum_{i=1}^n e^{(\beta_{10}+\beta_{20})A_i t}\left\{A_i - expit(\alpha_0^\top Z_i)\right\}\exp\left(\Lambda_{c0}(t)e^{\eta_0^\top D_i}\right)e^{\eta_0^\top D_i} dM_{ji}(t;\beta_{j0},\Lambda_j^*),$$

and

$$P_{3j}^{(a')}(t) := -\frac{1}{n}\sum_{i=1}^n e^{(\beta_{10}+\beta_{20})A_i t}\exp\left(\Lambda_{c0}(t)e^{\eta_0^\top D_i}\right) expit(\alpha_0^\top Z_i)e^{\alpha_0^\top Z_i}Z_i dM_{ji}(t;\beta_{j0},\Lambda_j^*),$$

where $D=[A,Z]^\top$. There exist, for $l=1,2,3$, some bounded $p_{lj}^{(a)}(t)$ and a neighborhood $\mathcal{B}$ of $\{\beta_0, S_{c0}(\cdot|\cdot,\cdot), \pi_0(\cdot), \Lambda^*(\cdot,\cdot)\}$ such that:

$$\sup_{t\in[0,\tau],\{\beta,S_c,\pi,\Lambda_j\}\in\mathcal{B}} \left|P_{lj}^{(a)}(t)-p_{lj}^{(a)}(t)\right| \xrightarrow{p} 0.$$

Assumption A* 2.

$$\sup_{Z\in\mathcal{Z}} |\Lambda_j^*(\tau, Z)| < \infty.$$

Assumption A* 3. *Let, for* $l = 0,1,2$

$$S_d^{(l)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) D_i^l e^{\eta_0^\top D_i}.$$

There exist, for $l = 01,2$*, some bounded* $s_d^{(l)}(t)$ *such that:*

$$\sup_{t\in[0,\tau]} \left| S_d^{(l)}(t) - s_d^{(l)}(t) \right| \xrightarrow{p} 0.$$

Assumption A* 4.

$$\int_0^\tau \left\{ \frac{s_d^{(2)}(t)}{s_d^{(0)}(t)} - \left(\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right)^2 \right\} E\left\{ Y(t) e^{\eta_0^\top D} \right\} d\Lambda_{c0}(t),$$

and

$$E\left[Z^\top Z \pi_0(Z_i) \left\{ 1 - \pi_0(Z_i) \right\} \right],$$

are positive definite.

Assumption B* 1. *Let, for* $l = 0,1$

$$P_1^{(b')}(t) := \frac{1}{n}\sum_{i=1}^{n} e^{(\beta_{10}+\beta_{20})A_i t} \left\{ S_c^*(t|A_i, Z_i) \right\}^{-1} \left\{ A_i - \pi^*(Z_i) \right\} Y_i(t) Z_i^l.$$

We assume that, for $l = 0,1$*, there exist* $p_l^{(b')}(t)$ *and a neighborhood* $\mathcal{B}$ *of*

$\{\beta_0, \Lambda_0(\cdot,\cdot), S_c^*(\cdot|\cdot,\cdot), \pi^*(\cdot)\}$ *such that :*

$$\sup_{t\in[0,\tau],\{\beta,\Lambda,S_c,\pi\}\in\mathcal{B}} \left| P_l^{(b')}(t) - p_l^{(b')}(t) \right| \xrightarrow{p} 0.$$

Assumption B* 2. *Let for $l = 0, 1$*

$$S_d^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) D_i,$$

$$S_z^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) Z_i,$$

and

$$S^{(0)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t),$$

where $D = [A, Z]^\top$.

We assume that, there exist $s_d^{(1)}(t), s_z^{(1)}(t), s^{(0)}(t)$ such that:

$$\sup_{t\in[0,\tau]} \left| S_d^{(1)}(t) - s_d^{(1)}(t) \right| \xrightarrow{p} 0,$$

$$\sup_{t\in[0,\tau]} \left| S_z^{(1)}(t) - s_z^{(1)}(t) \right| \xrightarrow{p} 0,$$

and

$$\sup_{t\in[0,\tau]}\left|S^{(0)}(t)-s^{(0)}(t)\right|\xrightarrow{p}0.$$

Assumption B* 3. *Let for $l=0,1$*

$$S^{(l)}_{wd}(t;S^*,\pi^*)=\frac{1}{n}\sum_{i=1}^{n}w_i(S^*,\pi^*)Y_i(t)D_i^l,$$

$$S^{(l)}_{wz}(t;S^*,\pi^*)=\frac{1}{n}\sum_{i=1}^{n}w_i(S^*,\pi^*)Y_i(t)Z_i^l.$$

We assume that, there exist $s^{(l)}_{wd}(t;S^*,\pi^*),s^{(l)}_{wz}(t;S^*,\pi^*)$ such that:

$$\sup_{t\in[0,\tau]}\left|S^{(l)}_{wd}(t;S^*,\pi^*)-s^{(l)}_{wd}(t;S^*,\pi^*)\right|\xrightarrow{p}0,$$

$$\sup_{t\in[0,\tau]}\left|S^{(l)}_{wz}(t;S^*,\pi^*)-s^{(l)}_{wz}(t;S^*,\pi^*)\right|\xrightarrow{p}0.$$

Assumption B* 4. $\int_0^\tau E\left[\left\{D-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2}Y(t)\right]dt$ *is positive definite.*

Assumptions for Theorem 7

Assumption A' 1. *There exist σ_6 such that:*

$$S_{2,n}(\beta_0,\hat{S}_c,\hat{\pi},\Lambda^*)-S_{2,n}(\beta_0,S_{c0},\pi_0,\Lambda^*)=\frac{1}{n}\sum_{i=1}^{n}\sigma_{6i}.$$

Assumption B' 1. *There exist* σ_7 *such that:*

$$S_{2,n}(\beta_0, S_c^*, \pi^*, \hat{\Lambda}) - S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda_0) = \frac{1}{n}\sum_{i=1}^{n}\sigma_{7i}.$$

Assumption C' 1. *Let*

$$h(t; A, Z) = \{A - \mathcal{E}_A(t; \beta_0, S_{c0}, \pi_0, Z)\},$$

$$P'(t) = \frac{1}{n}\sum_{i=1}^{n} h^2(t; A_i, Z_i) A_i Y_i(t),$$

and

$$Q_j'(t) = \frac{1}{n}\sum_{i=1}^{n} h^2(t; A_i, Z_i)\Lambda_{j0}(t, Z_i) Y_i(t).$$

We assume that, there exist $p'(t), q_j'(t)$ *and a neighborhood* $\mathcal{B}$ *of the true* $\{\beta_0, S_{c0}(\cdot|\cdot,\cdot), \pi_0(\cdot), \Lambda_0(\cdot,\cdot)\}$ *such that:*

$$\sup_{t\in[0,\tau],\{\beta,S_c,\pi,\Lambda\}\in\mathcal{B}} \left|P'(t) - p'(t)\right| \xrightarrow{p} 0,$$

and

$$\sup_{t\in[0,\tau],\{\beta,S_c,\pi,\Lambda\}\in\mathcal{B}} \left|Q_j'(t) - q_j'(t)\right| \xrightarrow{p} 0.$$

Assumption C' 2. *Let* $p'(t), q_j'(t)$ *as in assumption C'1, then for* $j = 1, 2,$ *we assume that:*

$$\int_0^\tau \left\{p'(u)\beta_j + q_j'(u)\right\} du > 0.$$

Remark 2. *Similarly to score 1, Assumption A' 1 can be proved assuming some regularity assumptions and assuming that there exist influence functions for* $\hat{S}_c(\cdot|\cdot,\cdot) - S_{c0}(\cdot|\cdot,\cdot)$ *and* $\hat{\pi}(\cdot) - \pi_0(\cdot)$. *In the same way Assumption B' 1 can be proved assuming some regularity assumptions and assuming that there exist an influence function for* $\hat{\Lambda}(\cdot,\cdot) - \Lambda_0(\cdot,\cdot)$.

3.8.4 Technical Quantities

We introduce here the technical quantities used in the paper.

Technical quantities of Theorem 3

We note that, by algebra:

$$
\begin{aligned}
& \{S_{1,n}\}_j(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) - \{S_{1,n}\}_j(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{e^{(\beta_1+\beta_2)A_i t} - e^{(\beta_{10}+\beta_{20})A_i t} + e^{(\beta_{10}+\beta_{20})A_i t}\right\}\hat{S}_c^{-1}(t \mid A_i, Z_i)\{A_i - \hat{\pi}(Z_i)\} \\
& \times \left\{dN_{ji}(t) - Y_i(t)\beta_j A_i dt - Y_i(t)d\hat{\Lambda}_j(t, Z_i; \beta)\right\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}\hat{S}_c^{-1}(t \mid A_i, Z_i)\{A_i - \hat{\pi}(Z_i)\} \\
& \times \left\{dN_{ji}(t) - Y_i(t)\beta_{j0} A_i dt - Y_i(t)d\hat{\Lambda}_j(t, Z_i; \beta_0)\right\}.
\end{aligned}
$$

By Lemma 12:

$$\begin{aligned}
= \quad & (\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_1^*+\beta_2^*)A_it}A_it\hat{S}_c^{-1}(t\mid A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\}dM_{ji}(t;\beta_j,\hat{\Lambda}) \\
& -(\beta_j-\beta_{j0})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_it}\hat{S}_c^{-1}(t\mid A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\}Y_i(t)A_idt \\
& -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_it}\hat{S}_c^{-1}(t\mid A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\}Y_i(t)d\left\{\hat{\Lambda}_j(t,Z_i;\beta)-\Lambda_j(t,Z_i;\beta_0)\right\},
\end{aligned}$$

where β_j^* is between β_j and β_0.

Therefore, for ease of reading we introduce the following additional notation, for $j=1,2$:

$$K^{(1)}(\beta,S_c,\pi) := \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_1+\beta_2)A_it}S_c^{-1}(t\mid A_i,Z_i)\left\{A_i-\pi(Z_i)\right\}Y_i(t)A_idt, \tag{3.34}$$

$$\begin{aligned}
K_j^{(2)}(\beta,S_c,\pi) \quad := \quad & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_it}S_c^{-1}(t\mid A_i,Z_i)\left\{A_i-\pi(Z_i)\right\}Y_i(t) \\
& \times d\left\{\hat{\Lambda}_j(t,Z_i;\beta)-\hat{\Lambda}_j(t,Z_i;\beta_0)\right\},
\end{aligned} \tag{3.35}$$

$$K_{ji}^{(3)}(t,\beta,S_c,\pi) := S_c^{-1}(t\mid A_i,Z_i)\left\{A_i-\pi(Z_i)\right\}dM_{ji}(t;\beta_j,\hat{\Lambda}), \tag{3.36}$$

$$\begin{aligned}
K_j^{(4)}(\beta,S_c,\pi,\Lambda) \quad = \quad & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_1+\beta_2)A_it}A_it\left\{S_c(t\mid A_i,Z_i)\right\}^{-1}\left\{A_i-\pi(Z_i)\right\}Y_i(t) \\
& \times d\left\{\Lambda_j(t,Z_i)-\Lambda_{j0}(t,Z_i)\right\}.
\end{aligned} \tag{3.37}$$

The introduction of $K_{ji}^{(3)}$ and $K_j^{(4)}$ will be clear to the reader in the proof of Lemma 7.

We now define all the technical quantities used in the Theorems.

Technical quantities of Theorem 4

$$\begin{aligned}\psi_i^{(a)} &= \left\{\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\{S_{c0}(t|A_i,Z_i)\}^{-1}\{A_i-\pi_0(Z_i)\}\,dM_{ji}(t;\beta_{j0},\Lambda_j^*)\right.\\ &\quad\left.+\int_0^\tau\left[\{p_{1j}^{(a)}(t)\}^\top\sigma_1 dt+\int_0^\tau p_{2j}^{(a)}(t)\sigma_2(t)dt-\int_0^\tau\{p_{3j}^{(a)}(t)\}^\top\sigma_3 dt\right]\right\}_{j=1,2}.\end{aligned}$$

$K^{(a)}$ is a 2X2 matrix with the following components:

$$K_{jj}^{(a)} = -K^{(1)}(\beta_0,S_{c0},\pi_0)-K_j^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*),$$

$$K_{12}^{(a)}=-K_1^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*),\quad K_{21}^{(a)}=-K_2^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*).$$

$$\begin{aligned}\psi_i^{(b)} &= \left[\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\{S_c^*(t|A_i,Z_i)\}^{-1}\{A_i-\pi^*(Z_i)\}\,dM_{ji}(t)\right.\\ &\quad\left.+\int_0^\tau\left\{\sigma_{4i}p_{1j}^{(b)}(t)dt+p_{2j}^{(b)}(t)d\sigma_{5i}(t)+dp_{2j}^{(b)}(t)\sigma_{5i}(t)\right\}\right]_{j=1,2}.\end{aligned}$$

$K^{(b)}$ is a 2X2 diagonal matrix with:

$$K_{jj}^{(b)} = -K^{(1)}(\beta_0,S^*,\pi^*)-K_j^{(2)}(\beta,S_c^*,\pi^*)/(\beta_j-\beta_{j0}).$$

Technical quantities of Corollary 1

$$\begin{aligned}
\psi_i^{(a')} &= \Bigg[\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_{c0}(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi_0(Z_i)\right\}dM_{ji}(t;\beta_{j0},\Lambda_j^*) \\
&\quad + \left(\int_0^\tau\left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)}-\left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right]s_d^{(0)}(t)d\Lambda_{c0}(t)\right)^{-1}\int_0^\tau\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}dM_i^c(t) \\
&\quad \times\left[\int_0^\tau\{p_{1j}^{(a')}\}^\top(t)dt-\int_0^\tau p_{2j}^{(a')}(t)\int_0^t d\Lambda_{c0}(u;\eta_0)\frac{s_d^{(1)}(u)}{s_d^{(0)}(u)}dudt\right] \\
&\quad + \int_0^\tau p_{2j}^{(a')}(t)\int_0^t\left\{s_d^{(0)}(u)\right\}^{-1}dM_i^c(u)dt \\
&\quad - \int_0^\tau\{p_{3j}^{(a')}\}^\top(t)\left(\mathrm{E}\left[Z^\top Z\pi_0(Z)\left\{1-\pi_0(Z)\right\}\right]\right)^{-1}Z_i\left\{A_i-\pi_0(Z_i)\right\}dt\Bigg]_{j=1,2}.
\end{aligned}$$

$$\begin{aligned}
\psi_i^{(b')} &= \Bigg[\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi^*(Z_i)\right\}dM_{ji}(t) \\
&\quad + \int_0^\tau\left(\left[\frac{1}{n}\sum_{i=1}^n\int_0^\tau\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2}Y_i(t)dt\right]^{-1}\int_0^\tau\left\{Z_i-\frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}\right\}dM_{ji}(t)\right)^\top \\
&\quad \times\left\{p_1^{(b')}(t)dt-p_0^{(b')}(t)\frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}dt\right\}+\int_0^\tau p_0^{(b')}(t)\left\{s_z^{(0)}(t)\right\}^{-1}dM_{ji}(t)\Bigg]_{j=1,2}.
\end{aligned}$$

$K^{(b')}$ is a 2X2 diagonal matrix with:

$$K_{jj}^{(b')} = \mathrm{E}\left[\int_0^\tau e^{(\beta_{10}+\beta_{20})At}\left\{S_c^*(t|A,Z)\right\}^{-1}\left\{A-\pi^*(Z)\right\}Y(t)\left[A-\frac{s_a^{(1)}(t)}{s^{(0)}(t)}\right]dt\right].$$

$$\begin{aligned}\psi_i^{(b'')} &= \left[\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t} \left\{S_c^*(t|A_i,Z_i)\right\}^{-1} \left\{A_i - \pi^*(Z_i)\right\} dM_{ji}(t)\right. \\ &\quad + \int_0^\tau \left(\left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right\}^{\otimes 2} Y_i(t)dt \right]^{-1} \int_0^\tau \left\{ Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)} \right\} dM_{ji}(t) \right)^\top \\ &\quad \left. \cdot \left\{ p_1^{(b')}(t)dt - p_0^{(b')}(t) \frac{s_{wz}^{(1)}(t;S_c^*,\pi^*)}{s_{wz}^{(0)}(t;S_c^*,\pi^*)} dt \right\} + \int_0^\tau p_0^{(b')}(t) \left\{ s_{wz}^{(0)}(t;S_c^*,\pi^*) \right\}^{-1} dM_{ji}(t) \right]_{j=1,2}.\end{aligned}$$

$K^{(b'')}$ is a 2X2 diagonal matrix with:

$$K_{jj}^{(b'')} = \mathrm{E}\left(\int_0^\tau e^{(\beta_{10}+\beta_{20})At} \left\{S_c^*(t|A,Z)\right\}^{-1} \left\{A - \pi^*(Z)\right\} Y(t) \left[A - \frac{s_{aw}^{(1)}(t;S^*,\pi^*)}{s_{aw}^{(0)}(t;S^*,\pi^*)} \right] dt \right).$$

Technical quantities of Theorem 7

We introduce the following notation:

$$J_{jj}^{(1)} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ A_i - \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i) \right\} A_i Y_i(t) dt,$$

$$J_{jj}^{(2)} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \partial_{\beta_j} \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i) Y_i(t) d\left\{ \Lambda_j^*(t,Z_i;\beta_{j0}) - \Lambda_{j0}(t,Z_i) \right\},$$

$$J_{jj}^{(3)} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ A_i - \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i) \right\} Y_i(t) \partial_{\beta_j} d\Lambda_j^*(t,Z_i),$$

$$J_{12}^{(1)} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \partial_{\beta_2} \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i) Y_i(t) d\left\{ \Lambda_1^*(t,Z_i;\beta_{10}) - \Lambda_{10}(t,Z_i) \right\},$$

$$J_{21}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_1}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)d\left\{\Lambda_2^*(t,Z_i;\beta_{10})-\Lambda_{20}(t,Z_i)\right\}.$$

$$\phi_i^{(a)} \quad = \quad \left[\int_0^\tau \{A_i - \mathcal{E}_{A_i}(t;\beta,S_c,\pi,Z_i)\}\,dM_{ji}(t;\beta_{j0},\Lambda_j^*) + \{\sigma_{6i}\}_j\right]_{j=1,2}.$$

$J^{(a)}$ is a 2X2 matrix with the following components $J_{jj}^{(a)} = J_{jj}^{(1)} + J_{jj}^{(2)}$ and $J_{12}^{(a)} = J_{12}^{(3)}$, $J_{21}^{(a)} = J_{21}^{(3)}$.

$$\phi_i^{(b)} \quad = \quad \left[\int_0^\tau \{A_i - \mathcal{E}_{A_i}(t;\beta,S_c,\pi,Z_i)\}\,dM_{ji}(t;\beta_{j0},\Lambda_j^*) + \{\sigma_{7i}\}_j\right]_{j=1,2}.$$

$J^{(b)}$ is a 2X2 diagonal matrix with $J_{jj}^{(b)} = J_{jj}^{(1)} + J_{jj}^{(3)}$.

3.8.5 Proofs of the main result

Proof of Theorem 2

Proof of Theorem 2. For $j = 1,\ldots,J$, we have:

$$\begin{aligned}
&\mathrm{E}\left[\{S_1\}_j(\beta_0;A,Z,S_c,\pi,\Lambda)\right] \\
&= \int_0^\tau \mathrm{E}\left[e^{\sum_{j=1}^J \beta_{j0}At} S_c^{-1}(t|A,Z)\{A-\pi(Z)\}\left\{dN_j(t) - Y(t)d\Lambda_j(t,Z) - Y(t)\beta_{j0}Adt\right\}\right] \\
&= \int_0^\tau \mathrm{E}[e^{\sum_{j=1}^J \beta_{j0}At} S_c^{-1}(t|A,Z)\{A-\pi(Z)\}\,dM_j(t)] \\
&\quad + \int_0^\tau \mathrm{E}[e^{\sum_{j=1}^J \beta_j At} S_c^{-1}(t|A,Z)\{A-\pi(Z)\}Y(t)d\left\{\Lambda_{j0}(t,Z) - \Lambda_j(t,Z)\right\}].
\end{aligned}$$

Therefore:

$$
\begin{aligned}
&\mathrm{E}\left[\{S_1\}_j(\beta_0;A,Z,S_c,\pi,\Lambda)\right] \\
&= \int_0^\tau \mathrm{E}(\mathrm{E}[e^{\sum_{j=1}^J \beta_{j0}At} S_c^{-1}(t|A,Z)\{A-\pi(Z)\}\mathrm{E}\{Y(t)|A,Z\}|Z] d\left\{\Lambda_{j0}(t,Z)-\Lambda_j(t,Z)\right\}) \\
&= \int_0^\tau \mathrm{E}(\mathrm{E}[S_c^{-1}(t|A,Z)S_{c0}(t|A,Z)\{A-\pi(Z)\}|Z] e^{-\sum_{l=1}^J \Lambda_{l0}(t,Z)} d\left\{\Lambda_{j0}(t,Z)-\Lambda_j(t,Z)\right\}) \\
&= \int_0^\tau \mathrm{E}\left(\left[S_c^{-1}(t|1,Z)S_{c0}(t|1,Z)\{1-\pi(Z)\}\pi_0(Z) - S_c^{-1}(t|0,Z)S_{c0}(t|0,Z)\pi(Z)\{1-\pi_0(Z)\}\right]\right. \\
&\quad \left. \times\, e^{-\sum_{l=1}^J \Lambda_{l0}(t,Z)} d\left\{\Lambda_{j0}(t,Z)-\Lambda_j(t,Z)\right\}\right),
\end{aligned}
$$

and the above is zero if either $\{S_c(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)\ \mathit{and}\ \pi(\cdot) = \pi_0(\cdot)\}$ or $\Lambda_j(\cdot,\cdot) = \Lambda_{j0}(\cdot,\cdot)$.

We have, for $j = 1,\ldots,J$:

$$
\begin{aligned}
&\mathrm{E}\left[\{S_2\}_j(\beta_0;A,Z,S_c,\pi,\Lambda)\right] \\
&\int_0^\tau \mathrm{E}\left[\{A-\mathcal{E}_A(t;\beta_0,S_c,\pi,Z)\}\left\{dN_j(t)-Y(t)d\Lambda_{j0}(t,Z)-Y(t)\beta_{j0}Adt\right\}\right] \\
&= \int_0^\tau \mathrm{E}\left[\{A-\mathcal{E}_A(t;\beta_0,S_c,\pi,Z)\}dM_j(t)\right] \\
&\quad + \int_0^\tau \mathrm{E}\left[\{A-\mathcal{E}_A(t;\beta_0,S_c,\pi,Z)\}Y(t)d\left\{\Lambda_{j0}(t,Z)-\Lambda_j(t,Z)\right\}\right] \\
&= \int_0^\tau \mathrm{E}\left(\mathrm{E}\left[\{A-\mathcal{E}_A(t;\beta_0,S_c,\pi,Z)\}\mathrm{E}\{Y(t)|A,Z\}|Z\right] d\left\{\Lambda_{j0}(t,Z)-\Lambda_j(t,Z)\right\}\right).
\end{aligned}
$$

Therefore

$$\begin{aligned}
&\mathrm{E}\left[\{S_2\}_j(\beta_0; A, Z, S_c, \pi, \Lambda)\right] \\
&= \int_0^\tau \mathrm{E}\Big(\mathrm{E}\Big[\{A - \mathcal{E}_A(t;\beta_0, S_c, \pi, Z)\}\, e^{-\sum_{l=1}^J \beta_{l0} A t} S_{c0}(t|A,Z)|Z\Big] \\
&\quad \times e^{-\sum_{l=1}^J \Lambda_{l0}(t,Z)} d\left\{\Lambda_{j0}(t,Z) - \Lambda_j(t,Z)\right\}\Big) \\
&= \int_0^\tau \mathrm{E}\Big\{\Big(e^{-\sum_{l=1}^J \beta_{l0} A t} S_{c0}(t|A=1,Z)\pi_0(Z) \\
&\quad -\mathcal{E}_A(t;\beta_0, S_c, \pi, Z)\Big[e^{-\sum_{l=1}^J \beta_{l0} A t} S_{c0}(t|A=1,Z)\pi_0(Z) + S_{c0}(t|A=0,Z)\{1-\pi_0(Z)\}\Big]\Big) \\
&\quad \times\, e^{-\sum_{l=1}^J \Lambda_{l0}(t,Z)} d\left\{\Lambda_{j0}(t,Z) - \Lambda_j(t,Z)\right\}\Big\},
\end{aligned}$$

and the above is zero if either $\{S_c(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)\ \mathit{and}\ \pi(\cdot) = \pi_0(\cdot)\}$ or $\Lambda_j(\cdot,\cdot) = \Lambda_{j0}(\cdot,\cdot)$. □

Proof of Theorems 3 and 4

Here we prove Theorems 3 and 4 that claim consistency and asymptotic normality of $\hat{\beta}^{(1)}$. We remind the reader that:

$$S_{1,n}(\beta; S_c, \pi, \Lambda) = \left\{\frac{1}{n}\sum_{i=1}^n \int_0^\tau e^{(\beta_1+\beta_2)A_i t} S_c^{-1}(t \mid A_i, Z_i)\{A_i - \pi(Z_i)\}\, dM_{ji}(t; \beta_j, \Lambda_j)\right\}_{j=1,2}.$$

By algebra we have the following decomposition of the score:

$$\begin{aligned}
S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) &= S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) - S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) \\
&\quad + S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) - S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*) \\
&\quad + S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*) - S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) \\
&\quad + S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*).
\end{aligned}$$

$S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)$, by Theorem 2, is sum of i.i.d mean zero terms. $S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) - S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda})$ can be written as $(\beta - \beta_0)$ times a positive definite matrix. $S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) - S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*)$ is negligible when the censoring model and the propensity score model are correctly specified, otherwise it is a sum of i.i.d mean zero terms plus a negligible term, as long as $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ and the rate of convergence of $\hat{\Lambda}(\cdot,\cdot)$ is $\sqrt{n}$.
$S_{1,n}(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*) - S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)$ is negligible when $\Lambda(\cdot,\cdot)$ is correctly specified, otherwise it is a sum of i.i.d mean zero terms plus a negligible term, as long as $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot), \pi^*(\cdot) = \pi_0(\cdot)$ and the rate of convergence of $\hat{S}_c(\cdot|\cdot,\cdot), \hat{\pi}(\cdot)$ is $\sqrt{n}$.

Therefore, in each scenario of Theorem 4, $\hat{\beta}^{(1)} - \beta_0$ can be written as a sum of i.i.d mean zero terms and hence the consistency and the asymptotic normality of $\hat{\beta}_1$.

The details of the above decomposition are contained in the following lemma:

Lemma 7. *For* β *in a compact neighborhood of* β_0*, under Assumptions S1-8 we have:*

$$\begin{aligned} S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) &= S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q^{(21)} + Q^{(3)} + K(\beta - \beta_0) \\ &\quad + O_p\left(n^{-1/2}|\beta_1 + \beta_2 - \beta_{10} - \beta_{20}| + |\beta_1 + \beta_2 - \beta_{10} - \beta_{20}|^2\right), \end{aligned}$$

where

$$\begin{aligned} Q_j^{(21)} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}\{S_c^*(t|A_i, Z_i)\}^{-1}\{A_i - \pi^*(Z_i)\}Y_i(t) \\ &\quad \times d\left\{\hat{\Lambda}_j(t, Z_i; \beta_0) - \Lambda_j^*(t, Z_i)\right\} = o_p(1), \end{aligned} \tag{3.38}$$

$$Q_j^{(3)} = \{S_{1,n}\}_j(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*) - \{S_{1,n}\}_j(\beta_0, S_c^*, \pi^*, \Lambda^*) = o_p(1), \tag{3.39}$$

and K is a 2x2 matrix with the following components:

$$K_{jj} \quad = \quad -K^{(1)}(\beta_0, S_c^*, \pi^*) - K_j^{(2)}(\beta, S_c^*, \pi^*)/(\beta_j - \beta_{j0}) - K_j^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*),$$

and

$$K_{12} = -K_1^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*), \quad K_{21} = -K_2^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*).$$

Moreover:

a) If $S_c^*(t|a,z) = S_c(t|a,z;\eta_0,\Lambda_{c0}) = S_{c0}(t|a,z)$, $\pi^*(z) = \pi(z;\alpha_0) = \pi_0(Z)$ and $\Lambda^*(\cdot,\cdot) \neq \Lambda_0(\cdot,\cdot)$, for some known functions S_c and π with $a_n = n^{-1/2}, b_n = n^{-1/2}$; specifically, under Assumptions A1-2: $\quad Q^{(21)} = o_p(n^{-1/2})$ and $Q^{(3)} = O_p(n^{-1/2})$.

b) If $\Lambda^*(t,z) = L(t,z;G_0,\gamma_0) = \Lambda_0(t,z)$, $S_c^*(\cdot|\cdot,\cdot) \neq S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) \neq \pi_0(\cdot)$, for some known function L with $c_n = n^{-1/2}$, specifically under Assumptions B1-2: $\quad Q^{(3)} = o_p(n^{-1/2})$ and $Q^{(21)} = O_p(n^{-1/2})$.

c) If $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot) = \pi_0(\cdot)$ and $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ with $a_n c_n = o(n^{-1/2})$ and $b_n c_n = o(n^{-1/2})$: $\quad Q^{(21)} = o_p(n^{-1/2})$ and $Q^{(3)} = o_p(n^{-1/2})$.

The proof of the Lemma is reported in Section 3.8.6.

We report in the following a detailed proof of Theorem 3 and 4.

Proof of Theorem 3. In Lemma 7 we prove that for β in a neighboorhood of β_0:

$$\begin{aligned} S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) \quad &= \quad S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + K(\beta - \beta_0) \\ &\quad + O_p\left(n^{-1/2}\left|\beta_1 + \beta_2 - \beta_{10} - \beta_{20}\right| + \left|\beta_1 + \beta_2 - \beta_{10} - \beta_{20}\right|^2\right) + o_p(1), \end{aligned}$$

where K is a 2x2 matrix with the following components:

$$K_{jj} = -K^{(1)}(\beta_0, S_c^*, \pi^*) - K_j^{(2)}(\beta_0, S_c^*, \pi^*)/(\beta_j - \beta_{j0}) - K_j^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*),$$

and

$$K_{12} = -K_4^{(1)}(\beta_0, S_c^*, \pi^*, \Lambda^*), \quad K_{21} = -K_4^{(2)}(\beta_0, S_c^*, \pi^*, \Lambda^*).$$

By double robustness of the score (Theorem 2), we have:

$$\mathrm{E}\left[S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)\right] = 0.$$

Therefore by Lemma 13, by Assumptions S1, S2 and S7 we have $S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) = O_p(n^{-1/2})$. Hence

$$\begin{aligned} S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) \quad &= \quad (\beta - \beta_0)K + o_p(1) \\ &\quad + O_p\left(n^{-1/2}\,|\beta_1 + \beta_2 - \beta_{10} - \beta_{20}| + |\beta_1 + \beta_2 - \beta_{10} - \beta_{20}|^2 + n^{-1/2}\right). \end{aligned} \tag{3.40}$$

By the above, we prove that, for $|\delta| < 1/2$:

$$S_{1,n}(\beta_0 \pm n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) \quad = \quad S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + n^{-\delta}K + O_p(n^{-1/2}).$$

If K is invertible we can conclude that either:

$$S_{1,n}(\beta_0 - n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) < 0 < S_{1,n}(\beta_0 + n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}),$$

or

$$S_{1,n}(\beta_0+n^{-\delta},\hat{S}_c,\hat{\pi},\hat{\Lambda}) < 0 < S_{1,n}(\beta_0-n^{-\delta},\hat{S}_c,\hat{\pi},\hat{\Lambda}).$$

Therefore by definition of $\hat{\beta}^{(1)}$, we can conclude that $\hat{\beta}^{(1)}-\beta_0 = O_p(n^{-\delta}) = o_p(1)$.

We are now left to prove that K is invertible. The latter simplifies according to which model is correctly specified. We therefore divide the proof into two cases.

- Case a): $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$ or $\hat{\Lambda}(\cdot,\cdot)$ does not depend on β or it depends on an initial estimator of it.

By Lemma 14, we have:

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^{n}\{A_i-\pi_0(Z_i)\}\{S_{c0}(t\mid A_i,Z_i)\}^{-1}Y_i(t)e^{(\beta_{10}+\beta_{20})A_it}\right| = O_p\left(n^{-1/2}\right).$$

By this and by Assumption S6, we have:

$$K_j^{(2)}(\beta,S_{c0},\pi_0) = O_p(n^{-1/2}|\beta_j-\beta_{j0}|) = o_p(1).$$

Therefore K simplifies and it has the following determinant:

$$\begin{aligned}
|K| &= \left\{K^{(1)}(\beta_0,S_{c0},\pi_0)\right\}^2 + K^{(1)}(\beta_0,S_{c0},\pi_0)K_2^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*)\\
&\quad +K_1^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*)K_2^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*) + K^{(1)}(\beta_0,S_{c0},\pi_0)K_1^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*)\\
&\quad -K_1^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*)K_2^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*)\\
&= K^{(1)}(\beta_0,S_{c0},\pi_0)\left\{K^{(1)}(\beta_0,S_{c0},\pi_0) + K_2^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*) + K_1^{(4)}(\beta_0,S_{c0},\pi_0,\Lambda^*)\right\}.
\end{aligned}$$

We prove now that both $K^{(1)}(\beta_0, S_{c0}, \pi_0)$ and $\left\{K^{(1)}(\beta_0, \pi_0, S_{c0}) + K_2^{(4)}(\beta_0, \pi_0, S_{c0}, \Lambda^*) + K_1^{(4)}(\beta_0, \pi_0, S_{c0}, \Lambda^*)\right\}$ are different from zero proving the invertibility of K.

We first focus on $K^{(1)}(\beta_0, S_{c0}, \pi_0)$.

By Assumptions S1 and S7, we have for some finite constant C:

$$\left|A_i\left\{A_i - \pi_0(Z_i)\right\}\int_0^\tau \left\{S_{c0}(t \mid A_i, Z_i)\right\}^{-1} e^{(\beta_{10}+\beta_{20})t} Y_i(t)dt\right| \leq C^{-1} e^{(\beta_{10}+\beta_{20})\tau}\tau < \infty. \tag{3.41}$$

Under model (3.1), $\mathrm{E}\left\{Y(t)|A,Z\right\} = S_{c0}(t \mid A, Z)e^{-(\beta_{10}+\beta_{20})At}e^{-\Lambda_{10}(t,Z)-\Lambda_{20}(t,Z)}$, we have:

$$\begin{aligned}
& \mathrm{E}\left[A\left\{A - \pi_0(Z)\right\}\int_0^\tau \left\{S_{c0}(t \mid A, Z)\right\}^{-1} e^{(\beta_{10}+\beta_{20})t} Y(t)dt\right] \\
= \; & \mathrm{E}\left(\mathrm{E}\left[A\left\{A - \pi_0(Z)\right\}\int_0^\tau \left\{S_{c0}(t \mid A, Z)\right\}^{-1} e^{(\beta_{10}+\beta_{20})At}\mathrm{E}\left\{Y(t)|A,Z\right\}dt\middle| Z\right]\right) \\
= \; & \mathrm{E}\left(\mathrm{E}\left[A\left\{A - \pi_0(Z)\right\}\int_0^\tau P\left\{T \geq t|A=0,Z\right\}dt\middle| Z\right]\right) \\
\geq \; & \mathrm{E}\left(\mathrm{E}\left[A\left\{A - \pi_0(Z)\right\}|Z\right]\int_0^\tau P\left\{T \geq t|A=0,Z\right\}dt\right) \\
\geq \; & \mathrm{E}\left[\mathrm{Var}\left(A|Z\right)\tau(1 - \mathrm{E}\left\{N(\tau)|A=0,Z\right\})\right].
\end{aligned}$$

Therefore, by Assumption S7 and S5, we have, for some positive ε:

$$\mathrm{E}\left[A\left\{A - \pi_0(Z)\right\}\int_0^\tau \left\{S_{c0}(t \mid A, Z)\right\}^{-1} e^{(\beta_{10}+\beta_{20})t} Y(t)dt\right] \; > \; \varepsilon > 0. \tag{3.42}$$

Hence, by Assumptions S1, S2, S7, by Hoeffding's inequality:

$$K^{(1)}(\beta_0, S_{c0}, \pi_0) = \mathrm{E}\left\{K^{(1)}(\beta_0, S_{c0}, \pi_0)\right\} + O_p(n^{-1/2}) > \varepsilon > 0.$$

We now focus on $K^{(1)}(\beta_0,\pi_0,S_{c0})+K_4^{(2)}(\beta_0,S_{c0},\pi_0,\Lambda^*)+K_4^{(1)}(\beta_0,S_{c0},\pi_0,\Lambda^*)$. Similarly, by Assumptions S5 and S8 we have:

$$\begin{aligned}
& \left|K^{(1)}(\beta_0,S_{c0},\pi_0)+K_4^{(2)}(\beta_0,S_{c0},\pi_0,\Lambda^*)+K_4^{(1)}(\beta_0,S_{c0},\pi_0,\Lambda^*)\right| \\
= & \left|\frac{1}{n}\sum_{i=1}^{n} A_i\{A_i-\pi_0(Z_i)\}\int_0^\tau \{S_{c0}(t\mid A_i,Z_i)\}^{-1}Y_i(t)e^{(\beta_{10}+\beta_{20})A_it}\right. \\
& \times[1+td\{\Lambda_1^*(t,Z_i)-\Lambda_{10}(t,Z_i)\}+td\{\Lambda_2^*(t,Z_i)-\Lambda_{20}(t,Z_i)\}]| \\
\geq & \left|\mathrm{E}\left(\{A-\pi_0(Z)\}\int_0^\tau \exp\{-\Lambda_{10}(t,Z)-\Lambda_{20}(t,Z)\}\right.\right. \\
& \times[A+Atd\{\Lambda_1^*(t,Z)-\Lambda_{10}(t,Z)\}+Atd\{\Lambda_2^*(t,Z)-\Lambda_{20}(t,Z)\}])|+O_p(n^{-1/2}) \\
\geq & \left|\mathrm{E}\left(\mathrm{E}[A\{A-\pi_0(Z)\}|Z]\int_0^\tau[1+td\{\Lambda_1^*(t,Z)-\Lambda_{10}(t,Z)\}+td\{\Lambda_2^*(t,Z)-\Lambda_{20}(t,Z)\}]\right)\right| \\
& \cdot(1-\mathrm{E}\{N(\tau)|A=0,Z\})+O_p(n^{-1/2}) \\
= & \mathrm{E}\left(\mathrm{Var}(A|Z)\int_0^\tau|1+td\{\Lambda_1^*(t,Z)-\Lambda_{10}(t,Z)\}+td\{\Lambda_2^*(t,Z)-\Lambda_{20}(t,Z)\}|\right. \\
& \times(1-\mathrm{E}\{N(\tau)|A=0,Z\}))+O_p(n^{-1/2})>\varepsilon.
\end{aligned}$$

We can therefore conclude that K is invertible.

- Case b): $\hat{\Lambda}(\cdot,\cdot)$ depends on the unknown β and $\Lambda^*(\cdot,\cdot)=\Lambda_0(\cdot,\cdot),S^*(\cdot|\cdot,\cdot)\neq S_{c0}(\cdot|\cdot,\cdot),\pi^*(\cdot)\neq\pi_0(\cdot)$.

By definition $K_j^{(4)}(\beta_0,S_c^*,\pi^*,\Lambda_0)=0$.

Again, we want to prove that K is invertible proving that the determinant is different from zero. Since $K_j^{(4)}(\beta_0,S_c^*,\pi^*,\Lambda_0)=0$, K is a diagonal matrix so we just need to verify that the diagonal elements are not null.

We have:

$$\begin{aligned}
\left|K_{jj}\right| &= \left|K^{(1)}(\beta_0, S_c^*, \pi^*) + K_j^{(2)}(\beta, S_c^*, \pi^*)/(\beta_j - \beta_{j0})\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t \mid A_i, Z_i)\right\}^{-1}\left\{A_i - \pi^*(Z_i)\right\}Y_i(t)\left\{A_i + \frac{1}{n}\sum_{l=1}^{n} q_{jl}(t)\right\}dt\right|,
\end{aligned}$$

where we call $q_{ji}(t)$ a function, such that:

$$\hat{\Lambda}_j(t, Z; \beta) - \hat{\Lambda}_j(t, Z; \beta_0) = (\beta_j - \beta_{j0}) * \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau q_{ji}(t).$$

Similarly to before, by Assumptions S1, S7, and Hoeffding's inequality, we have:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t \mid A_i, Z_i)\right\}^{-1}\left\{A_i - \pi^*(Z_i)\right\}Y_i(t)\left[A_i - \frac{1}{n}\sum_{l=1}^{n} q_{jl}(t)\right]dt \\
&= \mathrm{E}\left(\int_0^\tau \left\{S_c^*(t \mid A, Z)\right\}^{-1} S_{c0}(t \mid A, Z) e^{-\sum_{l=1}^{2}\Lambda_{l0}(t,Z)}\left\{A - \pi^*(Z)\right\}\left[A - \mathrm{E}(q_j(t))\right]dt\right) \\
&+ O_p(n^{-1/2}) \\
\geq\ & C\mathrm{E}\left(\int_0^\tau e^{-\sum_{l=1}^{2}\Lambda_{l0}(t,Z)}\mathrm{E}\left[\left\{A - \pi^*(Z)\right\}\left\{A - \mathrm{E}(q_j(t))\right\}|Z\right]dt\right) + O_p(n^{-1/2}).
\end{aligned}$$

Therefore, by Assumptions S7 and S8, we have $\left|K_{jj}\right| > \varepsilon + O_p(n^{-1/2})$ and hence K is invertible. □

Proof of Theorem 4. In Lemma 7 we prove that for β in a neighboorhood of β_0:

$$\begin{aligned}
S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) &= S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q^{(21)} + Q^{(3)} + K(\beta - \beta_0) \\
&+ O_p\left(n^{-1/2}\left|\beta_1 + \beta_2 - \beta_{10} - \beta_{20}\right| + \left|\beta_1 + \beta_2 - \beta_{10} - \beta_{20}\right|^2\right),
\end{aligned}$$

where

$$\begin{aligned} Q_j^{(21)} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t \mid A_i, Z_i)\right\}^{-1}\left\{A_i - \pi^*(Z_i)\right\} Y_i(t) \\ &\quad \times d\left\{\hat{\Lambda}_j(t, Z_i;\beta_0) - \Lambda_j^*(t, Z_i)\right\} = o_p(1), \end{aligned}$$

$$Q_j^{(3)} = \left\{S_{1,n}\right\}_j(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*) - \left\{S_{1,n}\right\}_j(\beta_0, S_c^*, \pi^*, \Lambda^*) = o_p(1).$$

In the proof of Theorem 3, we proved that $\hat{\beta} - \beta_0 = O_p(n^{-\delta})$ for any $|\delta| < 1/2$ and that K is invertible. Therefore we have:

$$\sqrt{n}(\hat{\beta}^{(1)} - \beta_0) = K^{-1}\left\{\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \sqrt{n}Q^{(21)} + \sqrt{n}Q^{(3)}\right\} + o_p(1). \quad (3.43)$$

We remind the reader that if the censoring model and the propensity score model are correctly specified, $Q^{(21)} = o_p(n^{-1/2})$. If $\Lambda(\cdot)$ is correctly specified, $Q^{(3)} = o_p(n^{-1/2})$. Hence, if every model is correctly specified, (3.43) simplifies and the asymptotic normality of $\hat{\beta}^{(1)}$ is obtained by the normality of $\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)$, that is a sum of i.i.d multivariate martingale integral.

If only the censoring model and the propensity score model are correctly specified, under Assumption A2, $Q^{(3)}$ is asymptotically linear. Asymptotic normality of $\hat{\beta}^{(1)}$ is therefore obtained by the normality of $\sqrt{n}S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \sqrt{n}Q^{(3)}$ that is a sum of i.i.d mean zero random variables.

If only the baseline hazard model is correctly specified, under Assumptions B2, $Q^{(21)}$ is asymptotically linear. Asymptotic normality of $\hat{\beta}^{(1)}$ is therefore obtained by the normality of $\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \sqrt{n}Q^{(21)}$that is a sum of i.i.d mean zero random variables.

In the following we prove the above statements in details.

- Case a):

We remind the reader that, if $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$, $Q^{(21)} = o_p(n^{-1/2})$ and $K_j^{(2)}(\beta, S_c^*, \pi^*) = o_p(1)$. In the proof of Theorem 3 we have proved that K simplifies to a 2x2 matrix with $K_{jj} = -K^{(1)}(\beta_0, S_c^*, \pi^*) - K_j^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*)$, and $K_{12} = -K_1^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*)$, $K_{21} = -K_2^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*)$.

Therefore, in (3.43) we are left with

$$\sqrt{n}(\hat{\beta} - \beta_0) = K^{-1}\sqrt{n}\left\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q^{(3)}\right\} + o_p(1). \tag{3.44}$$

Since $\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)$ is already a sum of i.i.d mean zero terms, with the help of Assumption A 2, we now prove that also term $Q^{(3)}$ can be written as the sum of i.i.d mean zero terms. We can then apply the multivariate central limit theorem to $\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \sqrt{n}Q^{(3)}$ and reach our conclusion.

We now look at the details. Using the fact that $\hat{\pi}(z) = \pi(z;\hat{\alpha})$, $\hat{S}_c(t|a,z) = S_c(t|a,z;\hat{\eta},\hat{\Lambda}_c)$ we have, by Taylor expansion:

$$\begin{aligned}
Q_j^{(3)} &= \sqrt{n}\left[\{S_{1,n}\}_j(\beta_0, \hat{S}_c, \hat{\pi}, \Lambda^*) - \{S_{1,n}\}_j(\beta_0, S_{c0}, \pi_0, \Lambda^*)\right] \\
&= \sqrt{n}\left[\{S_{1,n}\}_j(\beta_0, \hat{\eta}, \hat{\Lambda}_c, \hat{\alpha}, \Lambda^*) - \{S_{1,n}\}_j(\beta_0, \eta_0, \Lambda_{c0}, \alpha_0, \Lambda^*)\right] \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}\left\{D_{ij}^n(t, \eta_0, \Lambda_{c0}, \alpha_0, \Lambda^*)\right\}^{\top} \Delta dM_{ji}(t; \beta_{j0}, \Lambda_j^*) + o_p(n^{-1/2}),
\end{aligned}$$

where

$$D_{ij}^n(\eta, \Lambda_c, \alpha) := [\partial_{\eta} f(A_i, Z_i; \eta, \Lambda_c, \alpha), \partial_{\Lambda_c} f(A_i, Z_i; \eta, \Lambda_c, \alpha), \partial_{\alpha} f(A_i, Z_i; \eta, \Lambda_c, \alpha)]^{\top},$$

$$f(A_i, Z_i; \eta, \Lambda_c, \alpha) := \{A_i - \pi(Z; \hat{\alpha})\} S_c^{-1}(t|A, Z; \hat{\eta}, \hat{\Lambda}_c),$$

and

$$\Delta := [\hat{\eta} - \eta_0, \hat{\Lambda}_c(t) - \Lambda_{c0}(t), \hat{\alpha} - \alpha_0]^\top.$$

Standard algebra gives us:

$$\begin{aligned} D_{ij}^n(\eta, \Lambda_c, \alpha) &= [\{A_i - \pi(Z_i; \alpha)\} \partial_\eta S_c^{-1}(t|A, Z; \eta, \Lambda_c), \{A_i - \pi(Z_i; \alpha)\} \partial_{\Lambda_c} S_c^{-1}(t|A, Z; \eta, \Lambda_c) \\ &\quad , -\partial_\alpha \pi(Z_i; \alpha) S_c^{-1}(t|A, Z; \eta, \Lambda_c)]^\top. \end{aligned}$$

Moreover, by Assumption A2, we have:

$$\hat{\alpha} - \alpha_0 = O_p(n^{-1/2}), \quad \hat{\eta} - \eta_0 = O_p(n^{-1/2}), \quad \sup_{t \in [0,\tau]} \left\{\hat{\Lambda}_c(t) - \Lambda_{c0}(t)\right\} = O_p(n^{-1/2}).$$

Therefore, by the above and by Assumptions A1 and A2 we have:

$$\begin{aligned} Q_j^{(3)} &= \int_0^\tau \left[\left\{P_{1j}^{(a)}(t)\right\}^\top (\hat{\eta} - \eta_0) + P_{2j}^{(a)}(t) \left\{\hat{\Lambda}_c(t; \hat{\eta}) - \Lambda_{c0}(t)\right\} - \left\{P_{3j}^{(a)}(t)\right\}^\top (\hat{\alpha} - \alpha_0) \right] dt \\ &= \int_0^\tau \left[\left\{p_{1j}^{(a)}(t)\right\}^\top (\hat{\eta} - \eta_0) + p_{2j}^{(a)}(t) \left\{\hat{\Lambda}_c(t; \hat{\eta}) - \Lambda_{c0}(t)\right\} - \left\{p_{3j}^{(a)}(t)\right\}^\top (\hat{\alpha} - \alpha_0) \right] dt \\ &\quad + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[\left\{p_{1j}^{(a)}(t)\right\}^\top \sigma_{1i} + \int_0^\tau p_{2j}^{(a)}(t) \sigma_{2i}(t) - \int_0^\tau \left\{p_{3j}^{(a)}(t)\right\}^\top \sigma_{3i} \right] dt + o_p(n^{-1/2}). \end{aligned}$$

Therefore we have:

$$\sqrt{n}S_{1,n}(\beta_0, S_{c0}, \pi_0, \Lambda^*) + \sqrt{n}Q^{(3)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i^{(a)}(t) + o_p(1), \tag{3.45}$$

where

$$\begin{aligned}
\psi_{i,j}^{(a)} &= \int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\{S_{c0}(t \mid A_i, Z_i)\}^{-1}\{A_i - \pi_0(Z_i)\}\, dM_{ji}(t; \beta_{j0}, \Lambda_j^*) \\
&\quad + \int_0^\tau \left[\left\{p_{1j}^{(a)}(t)\right\}^\top \sigma_1 + p_{2j}^{(a)}(t)\sigma_2(t) - \int_0^\tau \left\{p_{3j}^{(a)}(t)\right\}^\top \sigma_3\right] dt.
\end{aligned}$$

By Theorem (2) and by construction of $Q^{(3)}$ the right hand side of (3.44) is a sum of i.i.d mean zero and the multivariate central limit theorem can be applied. Therefore, case a) of the Theorem is proven.

- Case b):

We remind the reader that, if $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$, $Q^{(3)} = o_p(n^{-1/2})$. In the proof of Theorem 3 we have proved that K simplifies to a 2x2 diagonal matrix with:

$$K_{jj} = -K^{(1)}(\beta_0, S_c^*, \pi^*) - K_j^{(2)}(\beta, S_c^*, \pi^*)/(\beta_j - \beta_{j0}).$$

Therefore, in (3.43) we are left with

$$\sqrt{n}(\hat{\beta} - \beta_0) = K^{-1}\sqrt{n}\left\{S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q^{(21)}\right\} + o_p(1). \tag{3.46}$$

Since $\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)$ is already a sum of i.i.d mean zero terms, with the help of Assumption B2, we now prove that also term $Q^{(21)}$ can be written as a sum of i.i.d mean zero terms.

We can then apply the multivariate central limit theorem to $\sqrt{n}S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \sqrt{n}Q^{(21)}$ and reach our conclusion.

We now look at the details.

By Assumption B2, we know that

$$\hat{\gamma}_j - \gamma_{j0} = O_p(n^{-1/2}), \quad \hat{G}_j(t) - G_{j0}(t) = O_p(n^{-1/2}).$$

Therefore by Taylor expansion we have:

$$\begin{aligned}
\sqrt{n}Q_j^{(21)} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t \mid A_i, Z_i)\right\}^{-1}\left\{A_i - \pi^*(Z_i)\right\}Y_i(t) \\
&\quad \times d\left\{L(t, Z; \hat{G}_j, \hat{\gamma}_j) - L(t, Z; G_{j0}, \gamma_{j0})\right\} + o_p(n^{-1/2}) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t \mid A_i, Z_i)\right\}^{-1}\left\{A_i - \pi^*(Z_i)\right\}Y_i(t) \\
&\quad \times \Big[(\hat{\gamma}_j - \gamma_{j0})^\top \partial_{\gamma_j} dL(t, Z; G_{j0}, \gamma_{j0}) + d\left\{\hat{G}_j(t) - G_{j0}(t)\right\}\partial_{\gamma_j} L(t, Z; G_{j0}, \gamma_{j0}) \\
&\quad + \left\{\hat{G}_j(t) - G_{j0}(t)\right\}\partial_{\gamma_j} dL(t, Z; G_{j0}, \gamma_{j0})\Big] + o_p(n^{-1/2}).
\end{aligned}$$

Hence, by the above and by Assumption B1 and B2, we have:

$$\begin{aligned}
Q_j^{(21)} &= \int_0^\tau \Big[(\hat{\gamma}_j - \gamma_{j0})^\top P_{1j}^{(b)}(t)dt + P_{2j}^{(b)}(t)d\left\{\hat{G}_j(t) - G_{j0}(t)\right\} \\
&\quad + dP_{2j}^{(b)}(t)\left\{\hat{G}_j(t) - G_{j0}(t)\right\}\Big] \\
&= \int_0^\tau \Big[(\hat{\gamma}_j - \gamma_{j0})^\top p_{1j}^{(b)}(t)dt + p_{2j}^{(b)}(t)d\left\{\hat{G}_j(t) - G_{j0}(t)\right\} \\
&\quad + dp_{2j}^{(b)}(t)\left\{\hat{G}_j(t) - G_{j0}(t)\right\}\Big] + o_p(n^{-1/2}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \Big[\sigma_{4i}p_{1j}^{(b)}(t)dt + p_{2j}^{(b)}(t)d\sigma_{5i}(t) + dp_{2j}^{(b)}(t)\sigma_{5i}(t)\Big] + o_p(n^{-1/2}).
\end{aligned} \tag{3.47}$$

Therefore we have:

$$\sqrt{n}S_{1,n}(\beta_0, S_{c0}, \pi_0, \Lambda^*) + \sqrt{n}Q^{(21)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i^{(b)}(t) + o_p(1) \tag{3.48}$$

where

$$\begin{aligned}\psi_{i,j}^{(b)} &= \int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\{S_c^*(t \mid A_i, Z_i)\}^{-1}\{A_i - \pi^*(Z_i)\}\, dM_{ji}(t; \beta_{j0}, \Lambda_j^*) \\ &\quad + \int_0^\tau \left[\sigma_{4i}p_{1j}^{(b)}(t)dt + p_{2j}^{(b)}(t)d\sigma_{5i}(t)dt + p_{2j}^{(b)}(t)\sigma_{5i}(t)\right].\end{aligned}$$

By Theorem 2 and by construction of $Q^{(21)}$ the right hand side of (3.46) is a sum of i.i.d mean zero random variable and the multivariate central limit theorem can be applied. Therefore, by the above together with (3.46), we can prove part b) of the Theorem.

- Case c):

We remind the reader that, if $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot) = \pi_0(\cdot)$ and $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$, we have $Q^{(21)} = o_p(n^{-1/2})$, $Q^{(3)} = o_p(n^{-1/2})$ and therefore the influence function in this case simplifies. In the proof of Theorem 3 we have proved that K simplifies to a 2x2 diagonal matrix with: $K_{jj} = -K^{(1)}(\beta_0, S_{c0}, \pi_0, \Lambda_0)$.

Indeed by this, by consistency of $\hat{\beta}^{(1)}$ proved in Lemma 3 and by (3.43), we have:

$$\sqrt{n}(\beta - \beta_0) \;=\; K^{-1}\sqrt{n}S_{1,n}(\beta_0, S_{c0}, \pi_0, \Lambda_0) + o_p(1).$$

We prove that $\sqrt{n}S_{1,n}(\beta_0, S_{c0}, \pi_0, \Lambda_0)$ is normal by martingale central limit theorem. Since here we assume that we plug in the true parameters, for ease of notation, in the following we will suppress the dependency of the martingale on β, Λ_0. We consider the following multivariate martingale: $M_i(t) = [M_{1i}(t), M_{2i}(t)]^\top$ with respect to the filtration $\mathcal{F}_t = \sigma\left\{N_j(s), Y(s+), A, Z \,:\, j = 1,2,\ 0 < s < t\right\}$. We consider the following two-dimensional vector: $M^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t h(u; A_i, Z_i) dM_i(u)$, where

$$h(t; A, Z) = e^{(\beta_{10}+\beta_{20})A_i t} \left\{S_{c0}(t \mid A, Z)\right\}^{-1} \left\{A - \pi_0(Z)\right\}.$$

Since $h(t; A, Z)$ is predictable with respect to the filtration, then $M^n(t)$ is a multivariate martingale too. By Assumption S3, we have

$$< M_{1i}(t), M_{2i}(t) > = < M_{1i}(t), M_{1j}(t) > = < M_{2i}(t), M_{2j}(t) > = < M_{1i}(t), M_{2j}(t) > = 0,$$

for each $i \neq j$ therefore:

$$\begin{aligned} < M_1^n(t), M_2^n(t) > &= < \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t h(u; A_i, Z_i) dM_{1i}(u), \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t h(u; A_i, Z_i) dM_{2i}(u) > \\ &= \frac{1}{n} \sum_{i,j=1}^n \int_0^t h^2(u; A_i, Z_i) d < M_{1i}(u), M_{2j}(u) > = 0, \end{aligned}$$

and so the two components of the multidimensional martingale $M^n(t)$ are orthogonal to each other. Therefore, we can apply the multidimensional version of the martingale central limit theorem of Rebolledo (Theorem 5 of Rebolledo (1978)).

First we verify Assumption 2 about the convergence of the variance. We have, by Assumption

S1, for $j = 1,2$:

$$\begin{aligned}
< M_j^n(t), M_j^n(t) > &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t h^2(u;A_i,Z_i)d\Lambda_j(u \mid A_i,Z_i)Y_i(u)\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t h^2(u;A_i,Z_i)\left\{d\Lambda_{j0}(u,Z_i)+\beta_{j0}Adu\right\}Y_i(u)\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t e^{2(\beta_{10}+\beta_{20})A_iu}\left\{S_{c0}(u \mid A_i,Z_i)\right\}^{-2}\left\{A_i-\pi_0(Z_i)\right\}^2\\
&\quad \times \left\{d\Lambda_{j0}(u,Z_i)+\beta_{j0}A_idu\right\}Y_i(u)\\
&= \int_0^t \left\{P(u)\beta_j+Q_j(u)\right\}du \xrightarrow{p} \int_0^t \left\{p(u)\beta_j+q_j(u)\right\}du = V_j(t),
\end{aligned}$$

and so Assumption 2 of the MCLT is verified.

We now look at Assumption 1 about the jumps of each component of the martingale. Rebolledo (1978) at pag. 39 claims that if the Lindeberg condition is verified, then Assumption 1 of its Theorem holds. We therefore needs to prove that, for any ε and any j:

$$\int_0^\tau \frac{1}{n}\sum_{i=1}^{n} h^2(u;A_i,Z_i)\mathbb{1}\left\{|h(u;A_i,Z_i)| > \sqrt{n}\varepsilon\right\}Y_i(t)\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\} \xrightarrow{p} 0,$$

by Assumption S1 and S7, we know that:

$$|h(t;A,Z)| \leq C_c^{-1}e^{(\beta_{10}+\beta_{20})\tau} < \infty,$$

so, we have:

$$\begin{aligned}
&\int_0^\tau \frac{1}{n}\sum_{i=1}^{n} h^2(u;A,Z)\mathbb{1}\left\{|h(u;A,Z)| > \sqrt{n}\varepsilon\right\}Y_i(t)\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\}\\
\leq &\int_0^\tau \frac{1}{n}\sum_{i=1}^{n} h^2(u;A,Z)\mathbb{1}\left\{C_c^{-1}\exp(\beta_{10}\tau+\beta_{20}\tau) > \sqrt{n}\varepsilon\right\}Y_i(t)\\
&\times\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\}.
\end{aligned}$$

Moreover, by Assumption S2, we also know that:

$$\begin{aligned}
&\left|\int_0^{\tau}\frac{1}{n}\sum_{i=1}^{n}h^2(u;A,Z)\mathbb{1}\left\{C_c^{-1}e^{(\beta_{10}+\beta_{20})\tau}>\sqrt{n}\varepsilon\right\}Y_i(t)\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_i dt\right\}\right| \\
\leq\ & C_c^{-2}e^{2(\beta_{10}+\beta_{20})\tau}\mathbb{1}\left\{C_c^{-1}e^{(\beta_{10}+\beta_{20})\tau}>\sqrt{n}\varepsilon\right\}\tau\left|L_j+\beta_{j0}\right|\xrightarrow{p}0,
\end{aligned}$$

and so Assumption 1 of the martingale central limit theorem holds.

Therefore, we can conclude that

$$\sqrt{n}S_{2,n}(\beta_0,S_{c0},\pi_0,\Lambda_0)=M^n(t)\xrightarrow{D}\mathcal{N}(0,V(\tau)).$$

Therefore part c) of the Theorem is proven. □

Proof of Theorem 5

Proof of Theorem 5. We prove separately that $\hat{W}_{jj}^{(c)}-W_{jj}^{(c)}=o_p(1)$ and that $\hat{V}_{jj}^{(c)}(\tau)-V_{jj}^{(c)}(\tau)=o_p(1)$.

We have:

$$
\begin{aligned}
&\hat{W}_{jj}^{(c)} - W_{jj}^{(c)} \\
&= \frac{1}{n}\sum_{i=1}^{n} A_i\{A_i - \hat{\pi}(Z_i)\}\int_0^{X_i}\left\{\hat{S}_c(t \mid A_i, Z_i)\right\}^{-1} e^{(\hat{\beta}_1+\hat{\beta}_2)t}dt \\
&- \mathrm{E}\left[A\{A-\pi_0(Z)\}\int_0^{X}\{S_{c0}(t \mid A, Z)\}^{-1} e^{(\beta_{10}+\beta_{20})t}dt\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} A_i\{A_i - \pi_0(Z_i)\}\int_0^{X_i}\{S_{c0}(t \mid A_i, Z_i)\}^{-1} e^{(\beta_{10}+\beta_{20})t}dt \\
&- \mathrm{E}\left[A\{A-\pi_0(Z)\}\int_0^{X}\{S_{c0}(t \mid A, Z)\}^{-1} e^{(\beta_{10}+\beta_{20})t}dt\right] \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^{X_i} A_i\{A_i - \pi_0(Z_i)\}\{S_{c0}(t \mid A_i, Z_i)\}^{-1}\left\{e^{(\hat{\beta}_1+\hat{\beta}_2)t} - e^{(\beta_{10}+\beta_{20})t}\right\}dt \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^{X_i} A_i\left[\{A_i - \hat{\pi}(Z_i)\}\left\{\hat{S}_c(t \mid A_i, Z_i)\right\}^{-1} - \{A_i - \pi_0(Z_i)\}\{S_{c0}(t \mid A_i, Z_i)\}^{-1}\right] \\
&\times \left\{e^{(\hat{\beta}_1+\hat{\beta}_2)t} - e^{(\beta_{10}+\beta_{20})t}\right\}dt \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^{X_i} A_i\left[\{A_i - \hat{\pi}(Z_i)\}\left\{\hat{S}_c(t \mid A_i, Z_i)\right\}^{-1} - \{A_i - \pi_0(Z_i)\}\{S_{c0}(t \mid A_i, Z_i)\}^{-1}\right] \\
&\times e^{(\beta_{10}+\beta_{20})t}dt \\
&= Q_1 - Q_2 + Q_3 + Q_4 + Q_5.
\end{aligned}
$$

As before, by Assumptions S1 and S7, by Hoeffding's inequality we have: $Q_1 - Q_2 = O_p(n^{-1/2})$. by Assumptions S1 and S7 and by Lemma 12, we get:

$$
|Q_3| \le \tau C_c^{-1}\left\{e^{(\hat{\beta}_1+\hat{\beta}_2)\tau} - e^{(\beta_{10}+\beta_{20})\tau}\right\} = \tau C_c^{-1} e^{(\beta_1^*+\beta_2^*)\tau}\tau\left(\hat{\beta}_1 + \hat{\beta}_2 - \beta_{10} - \beta_{20}\right).
$$

where β_j^* are points between $\hat{\beta}_j$ and β_{j0}. Therefore, by consistency of the estimator $\hat{\beta}$, we have $|Q_3| = o_p(1)$.

By Assumption 1 and by consistency of β we have $Q_4 = o_p(1)$. By this and by Assumptions 1, S1, S7, we have:

$$\begin{aligned}
|Q_5| &\leq \sup_{t\in[0,\tau],Z\in\mathcal{Z},A\in 0,1} \left\{\left|\hat{S}_c^{-1}(t \mid A,Z) - \{S_{c0}(t \mid A,Z)\}^{-1}\right| + \left|\hat{S}_c^{-1}(t \mid A,Z)\right| \left|\hat{\pi}(Z) - \pi_0(Z)\right|\right\} \\
&\quad \times \tau e^{(\beta_{10}+\beta_{20})\tau} \\
&= o_p(1). \qquad (3.49)
\end{aligned}$$

We can therefore conclude that $\hat{W} - W = o_p(1)$.

We have:

$$\begin{aligned}
&\hat{V}_{jj}^{(c)}(\tau) - V_{jj}^{(c)}(\tau) \\
= \; &\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{2(\hat{\beta}_1+\hat{\beta}_2)A_iX_i}\hat{S}_c^{-2}(X_i \mid A_i,Z_i)\left\{A_i - \hat{\pi}(Z_i)\right\}^2 dN_{ji}(t) - \int_0^{\tau}\left\{p(t)\beta_j + q_j(t)\right\}dt \\
= \; &\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{2(\beta_{10}+\beta_{20})A_it}\left\{S_{c0}(X_i \mid A_i,Z_i)\right\}^{-2}\left\{A_i - \pi_0(Z_i)\right\}^2 dM_{ji}(t) \\
&+\int_0^{\tau}\left\{P(t)\beta_j + Q_j(t) - p(t)\beta_j - q_j(t)\right\}dt \\
&+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{S_{c0}(X_i \mid A_i,Z_i)\right\}^{-2}\left\{A_i - \pi_0(Z_i)\right\}^2\left\{e^{2(\hat{\beta}_1+\hat{\beta}_2)t} - e^{2(\beta_{10}+\beta_{20})t}\right\}dN_{ji}(t) \\
&+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\hat{S}_c^{-2}(X_i \mid A_i,Z_i)\left\{A_i - \hat{\pi}(Z_i)\right\}^2 - \left\{S_{c0}(X_i \mid A_i,Z_i)\right\}^{-2}\left\{A_i - \pi_0(Z_i)\right\}^2\right] \\
&\times\left\{e^{2(\hat{\beta}_1+\hat{\beta}_2)t} - e^{2(\beta_{10}+\beta_{20})t}\right\}dN_{ji}(t) \\
&+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\hat{S}_c^{-2}(X_i \mid A_i,Z_i)\left\{A_i - \hat{\pi}(Z_i)\right\}^2 - \left\{S_{c0}(t \mid A_i,Z_i)\right\}^{-2}\left\{A_i - \pi_0(Z_i)\right\}^2\right] \\
&\times e^{2(\beta_{10}+\beta_{20})t}dN_{ji}(t) \\
= \; &E_1 + E_2 + E_3 + E_4 + E_5.
\end{aligned}$$

For E_1 we notice that, by Assumptions S1 and S7, we have:

$$e^{2(\beta_{10}+\beta_{20})A_iX_i}\{S_{c0}(X_i \mid A_i, Z_i)\}^{-2}\{A_i - \pi_0(Z_i)\}^2 \leq C_c^{-2}e^{2(\beta_{10}+\beta_{20})A_i\tau} < \infty,$$

and so, by Lemma 13, we have $E_1 = o_p(1)$.
By Assumption S1, we can prove that $E_2 = o_p(1)$.
Similarly to what we have done for Q_3, Q_4 and Q_5, we can prove that $E_3 = o_p(1)$, $E_4 = o_p(1)$ and $E_5 = o_p(1)$.
Therefore $\hat{V}_j - V_j = o_p(1)$. □

Proof of Corollary 1

The proof of Corollary 1 follows directly from finding the influence functions defined in Assumption A 2 and Assumption B 2 for the specific working models $S_c(t|a,z;\eta,\Lambda_c) = \exp\left(-\Lambda_c e^{\eta^\top d}\right)$, where $D = [A,Z]^\top$ and $\pi(z;\alpha) = \left\{1+\exp(-\alpha^\top z)\right\}^{-1}$ for case a) and $\Lambda_j(t,z;G_j,\gamma_j) = G_j(t) + \gamma_j^\top zt$ for case b) respectively. We indeed remind the reader that under case a), we have:

$$\sqrt{n}(\hat{\beta} - \beta_0) = K^{-1}\sqrt{n}\left\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q^{(3)}\right\} + o_p(1),$$

where $Q^{(3)}$, defined in Lemma 7, directly depends on the form of the influence functions of estimators $\hat{\alpha}, \hat{\Lambda}_c, \hat{\eta}$. Under case b), we instead have:

$$\sqrt{n}(\hat{\beta} - \beta_0) = K^{-1}\sqrt{n}\left\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q^{(21)}\right\} + o_p(1),$$

where $Q^{(21)}$, defined in Lemma 7, directly depends on the form of the influence functions of estimators $\hat{G}, \hat{\gamma}$.

The next Lemma defines the specific form of $Q^{(3)}$ when the logistic model and the Cox model are assumed on the propensity score and the censoring distribution, respectively.

Lemma 8. *We assume,* $\pi(Z;\alpha) = \left\{1+\exp(-\alpha^\top Z)\right\}^{-1}$ *and* $S_c(t|A,Z;\eta,\Lambda_c) = \exp\left(-\Lambda_c e^{\eta^\top D}\right)$. *Under Assumptions S2, S4 and Assumption A*1- A*4 we have:*

$$
\begin{aligned}
\sqrt{n} Q_j^{(3)} \quad = \quad & \left(\int_0^\tau \left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)} - \left\{ \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right\}^2 \right] s_d^{(0)}(t) d\Lambda_{c0}(t) \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau \left\{ D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right\} dM_i^c(t) \\
& \times \left[\int_0^\tau \left\{ p_1^{(a')} \right\}^\top (t) dt - \int_0^\tau p_2^{(a')}(t) \int_0^t d\Lambda_{c0}(u;\eta_0) \frac{s_d^{(1)}(u)}{s_d^{(0)}(u)} dt \right] \\
& + \int_0^\tau p_2^{(a')}(t) \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t \left\{ s_d^{(0)}(u) \right\}^{-1} dM_i^c(u) \\
& - \int_0^\tau \left\{ p_3^{(a')} \right\}^\top (t) \left(E\left[Z^\top Z \pi_0(Z_i) \left\{ 1 - \pi_0(Z_i) \right\} \right] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \left\{ A_i - \pi_0(Z_i) \right\} dt \\
& + o_p(1).
\end{aligned}
$$

The next lemma defines the specific form of $Q^{(21)}$ when the traditional additive hazard model is assumed on the cause-specific hazards.

Lemma 9. *Let* $\Lambda_j(t,Z;G_j,\gamma_j) = G_j(t) + \gamma_j^\top Zt$ *and let* γ_j *be estimated by* (3.10) *in the main document and* $G_j(t)$ *be estimated using* (3.11) *in the main document. Under Assumptions S2, S4 and*

*Assumptions B*2 and B*4 it holds:*

$$
\begin{aligned}
\sqrt{n}Q_j^{(21)} &= \int_0^\tau \left(\left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right\}^{\otimes 2} Y_i(t)dt \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau \left\{ Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)} \right\} dM_{ji}(t) \right)^\top \\
&\quad \times \left\{ p_1^{(b')}(t)dt - p_0^{(b')}(t) \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)} dt \right\} \\
&\quad + \int_0^\tau p_0^{(b')}(t) \left\{ s_z^{(0)}(t) \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n dM_{ji}(t) + o_p(1).
\end{aligned}
$$

Lemma 10. *Let* $\Lambda_j(t,Z;G_j,\gamma_j) = G_j(t) + \gamma_j^\top Zt$ *and let* γ_j *be estimated by* (3.10) *in the main document and* $G_j(t)$ *be estimated using* (3.12) *in the main document. Under Assumptions S2, S4 and Assumptions B*3 and B*4 it holds:*

$$
\begin{aligned}
\sqrt{n}Q_j^{(21)} &= \int_0^\tau \left(\left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right\}^{\otimes 2} Y_i(t)dt \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau \left\{ Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)} \right\} dM_{ji}(t) \right)^\top \\
&\quad \times \left\{ p_1^{(b')}(t)dt - p_0^{(b')}(t) \frac{s_{wz}^{(1)}(t;S_c^*,\pi^*)}{s_{wz}^{(0)}(t;S_c^*,\pi^*)} dt \right\} \\
&\quad + \int_0^\tau p_0^{(b')}(t) \left\{ s_{wz}^{(0)}(t;S_c^*,\pi^*) \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n dM_{ji}(t) + o_p(1).
\end{aligned}
$$

Theorem 4 together with Lemma 8 proves part a) of the corollary. Theorem 4 together with Lemma 9 proves part b1) of the corollary. Theorem 4 together with Lemma 10 proves part b2) of the corollary.

Proof of Theorems 6, 7 and 8

Proofs of Theorems 6, 7 and 8 use similar ideas and techniques used in the proofs of Theorems 3, 7 and 8. We therefore report here a sketch of their proofs.

Proof of Theorem 6 (sketch). Under Assumptions S1, it follows that:

$$S_{2,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) = S_{2,n}(\beta, S_c^*, \pi^*, \Lambda^*) + o_p(1).$$

By Taylor expansion we have:

$$S_{2,n}(\beta, S_c^*, \pi^*, \Lambda^*) = S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \nabla_\beta S_{2,n}(\beta^*, S_c^*, \pi^*, \Lambda^*)\,(\beta - \beta_0)^\top,$$

where β^* lies between β and β_0.

By double robustness of the score (Theorem 2), and by application of Hoeffding's inequality under Assumptions S2 and S7, we have: $S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) = E\left\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)\right\} + O_p(n^{-1/2}) = O_p(n^{-1/2})$. Therefore we have:

$$S_{2,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) = \nabla_\beta S_{2,n}(\beta^*, S_c^*, \pi^*, \Lambda^*)\,(\beta - \beta_0)^\top + O_p(n^{-1/2}). \tag{3.50}$$

We now focus on $\nabla := \nabla_\beta S_{2,n}(\beta^*, S_c^*, \pi^*, \Lambda^*)$. We have as diagonal element, for $j = 1, 2$:

$$\begin{aligned}
\nabla_{jj} &= \partial_{\beta_j}\left\{S_{2,n}\right\}_j(\beta^*, S_c^*, \pi^*, \Lambda^*) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)\,dM_{ji}(t;\beta_j^*, \Lambda_j^*) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{A_i - \mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)\right\} A_i Y_i(t) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{A_i - \mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)\right\} Y_i(t)\,\partial_\beta d\Lambda^*(t, Z_i;\beta^*) \\
&= Q_1 + Q_2 + Q_3.
\end{aligned}$$

We have:

$$\begin{aligned}
Q_1 &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)dM_{ji}(t)\\
&\quad +(\beta_j^*-\beta_{j0})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)A_iY_i(t)dt\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)d\left\{\Lambda_j^*(t,Z_i;\beta_j^*)-\Lambda_j^*(t,Z_i;\beta_{j0})\right\}\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)d\left\{\Lambda_j^*(t,Z_i;\beta_{j0})-\Lambda_{j0}(t,Z_i)\right\}\\
&= Q_{11}+Q_{12}+Q_{13}+Q_{14}.
\end{aligned}$$

Q_{11} is a martingale integral with bounded integrand by Assumption S7. Therefore, by concentration inequality of martingale integral is $o_p(1)$. Under Assumption S6, we have

$$Q_{13} = (\beta_j^*-\beta_{j0})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \partial_{\beta_j}\mathcal{E}_A(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)\mathrm{E}(q_j(t))dt,$$

where we call $q_{ji}(t)$ a function, such that:

$$\hat{\Lambda}_j(t,Z;\beta)-\hat{\Lambda}_j(t,Z;\beta_0) \quad = \quad (\beta_j-\beta_{j0})*\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau q_{ji}(t).$$

Term $Q_{14}=0$ if $\Lambda^*=\Lambda_0$. What about Q_3? If $S_c^*=S_{c0}$ and $\pi^*=\pi_0$, because everything is bounded, by Assumption S7 we have something along the following line:

$$
\begin{aligned}
Q_3 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{A_i - \mathcal{E}_{A_i}(t;\beta,S_{c0},\pi_0,Z_i)\}Y_i(t)\partial_\beta d\Lambda(t,Z_i;\beta^*) \\
&= \int_0^{\tau}\mathrm{E}\left[\left\{A - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{J}\beta_j At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{J}\beta_j At}S_{c0}(t|A,Z)|Z\right]}\right\}Y(t)\partial_\beta d\Lambda(t,Z;\beta^*)\right] + o_p(1) \\
&= \int_0^{\tau}\mathrm{E}\left[\left\{A - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{J}\beta_j At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{J}\beta_j At}S_{c0}(t|A,Z)|Z\right]}\right\}\mathrm{E}\{Y(t)|A,Z\}\partial_\beta d\Lambda(t,Z;\beta^*)\right] + o_p(1) \\
&= \int_0^{\tau}\mathrm{E}\left[\left\{A - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{J}\beta_j At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{J}\beta_j At}S_{c0}(t|A,Z)|Z\right]}\right\}\right. \\
&\qquad \left.\times e^{-\sum_{j=1}^{J}\beta_j At}e^{-\sum_{j=1}^{J}\Lambda_j(t,Z)}S_{c0}(t|A,Z)\partial_\beta d\Lambda(t,Z;\beta^*)\right] + o_p(1) \\
&= o_p(1).
\end{aligned}
$$

Therefore we have:

$$
\nabla_{jj} = \partial_{\beta_j}\{S_{2,n}\}_j(\beta^*,S_c^*,\pi^*,\Lambda^*) = (\beta_j^* - \beta_{j0})(J_{jj}^{(1')} + J_{jj}^{(2')}) + J_{jj}^{(1)} + J_{jj}^{(2)} + J_{jj}^{(3)},
$$

where

$$
J_{jj}^{(1')} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)A_iY_i(t)dt,
$$

$$
J_{jj}^{(2')} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)\mathrm{E}(q_j(t))dt,
$$

$$J_{jj}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{A_i - \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)\}A_iY_i(t),$$

$$J_{jj}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_j}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)d\left\{\Lambda_j^*(t,Z_i;\beta_{j0})-\Lambda_{j0}(t,Z_i)\right\},$$

$$J_{jj}^{(3)} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{A_i - \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)\}Y_i(t)\partial_{\beta_j}d\Lambda_j^*(t,Z_i,\beta_j^*).$$

We notice that $J_{jj}^{(2)} = 0$ if $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ and $J_{jj}^{(3)} = o_p(1)$ if or $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$ or $\hat{\Lambda}(\cdot,\cdot)$ does not depend on the unknown β.

On the other hand, similarly, we have:

$$\begin{aligned}
\nabla_{12} &= \partial_{\beta_2}\{S_{2,n}\}_1(\beta^*,S_c^*,\pi^*,\Lambda^*)\\
&= (\beta_1^*-\beta_{10})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_2}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)A_iY_i(t)dt\\
&\quad +(\beta_1^*-\beta_{10})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_2}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)\mathrm{E}(q_j(t))dt\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_2}\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)Y_i(t)d\{\Lambda_1^*(t,Z_i;\beta_{10})-\Lambda_{10}(t,Z_i)\}+o_p(1)\\
&= (\beta_1^*-\beta_{10})(J_{12}^{(1')}+J_{12}^{(2')})+J_{12}^{(1)},
\end{aligned}$$

and

$$\begin{aligned}
\nabla_{21} &= \partial_{\beta_1}\{S_{2,n}\}_2(\beta^*, S_c^*, \pi^*, \Lambda^*) \\
&= (\beta_2^* - \beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_1}\mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)A_iY_i(t)dt \\
&\quad +(\beta_2^* - \beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_1}\mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)Y_i(t)\mathrm{E}(q_j(t))dt \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\partial_{\beta_1}\mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)Y_i(t)d\left\{\Lambda_2^*(t, Z_i;\beta_{20}) - \Lambda_{20}(t, Z_i)\right\} \\
&= (\beta_2^* - \beta_{20})(J_{21}^{(1')} + J_{21}^{(2')}) + J_{21}^{(1)},
\end{aligned}$$

where the last terms $J_{12}^{(1)} = J_{21}^{(1)} = 0$ if $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$.

Therefore we have:

$$\nabla = J + \begin{bmatrix} \beta_1^* - \beta_{10} \\ \beta_2^* - \beta_{20} \end{bmatrix} J'$$

where the above multiplication is intended componentwise and

$$J = \begin{bmatrix} J_{11}^{(1)} + J_{11}^{(2)} + J_{11}^{(3)} & J_{12}^{(1)} \\ J_{21}^{(1)} & J_{22}^{(1)} + J_{22}^{(2)} + J_{22}^{(3)} \end{bmatrix},$$

and

$$J' = \begin{bmatrix} J_{11}^{(1')} + J_{11}^{(2')} & J_{12}^{(1')} + J_{12}^{(2')} \\ J_{21}^{(1')} + J_{21}^{(2')} & J_{22}^{(1')} + J_{22}^{(2')} \end{bmatrix}.$$

We will prove that J is invertible. If this is the case, for any $|\delta| < 1/2$, by (3.50) and the above we have:

$$S_{2,n}(\beta_0 \pm n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) \;=\; n^{-\delta} J + n^{-2\delta} J' + O_p(n^{-1/2}).$$

We can therefore conclude that either:

$$S_{2,n}(\beta_0 - n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) < 0 < S_{2,n}(\beta_0 + n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}),$$

or

$$S_{2,n}(\beta_0 + n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) < 0 < S_{2,n}(\beta_0 - n^{-\delta}, \hat{S}_c, \hat{\pi}, \hat{\Lambda}).$$

Therefore by definition of $\hat{\beta}$, we can conclude that $\hat{\beta} - \beta_0 = O_p(n^{-\delta})$.

We now prove that J is invertible proving that its determinant is different from zero. J simplifies accordingly to which model is correct. We therefore divide the proof of its invertibility in two cases.

- Case a): J is invertible if $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$.

Noticing that $\partial_{\beta_1} \mathcal{E}_A(t; \beta^*, S_c^*, \pi^*, Z) = \partial_{\beta_2} \mathcal{E}_A(t; \beta^*, S_c^*, \pi^*, Z)$ and $J_{11}^{(1)} = J_{22}^{(1)}$, after some algebra we have:

$$\begin{aligned} |J| &= (J_{11}^{(1)} + J_{11}^{(2)})(J_{22}^{(1)} + J_{22}^{(2)}) - J_{12}^{(1)} J_{21}^{(1)} + o_p(1) \\ &= J_{11}^{(1)}(J_{11}^{(1)} + J_{11}^{(2)} + J_{22}^{(2)}) + o_p(1). \end{aligned}$$

We now prove that both $J_{11}^{(1)} \neq 0$ and $J_{11}^{(1)} + J_{11}^{(2)} + J_{22}^{(2)} \neq 0$. Those would prove that $|J| \neq 0$ and so that J is invertible.

Under model (3.1), $\mathrm{E}\{Y(t)|A,Z\} = Sc0c(t|A,Z)e^{-(\beta_{10}+\beta_{20})At}e^{-\Lambda_{10}(t,Z)-\Lambda_{20}(t,Z)}$, therefore we have:

$$
\begin{aligned}
& \mathrm{E}\left[\left\{A - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}AY(t)dt\right] \\
= & \int_0^\tau \mathrm{E}\left[\left\{A - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}\mathrm{E}\{Y(t)|A,Z\}Adt\right],
\end{aligned}
$$

and so

$$
\begin{aligned}
= & \int_0^\tau \mathrm{E}\left[\left\{A - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}e^{-\sum_{j=1}^{2}\beta_{j0}At}e^{-\sum_{j=1}^{2}\Lambda_{j0}(t,Z)}S_{c0}(t|A,Z)Adt\right] \\
= & \int_0^\tau \mathrm{E}\left[\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]\left\{1 - \frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}\right. \\
& \left.\times e^{-\sum_{j=1}^{2}\Lambda_{j0}(t,Z)}dt\right].
\end{aligned}
$$

The above is strictly different from zero under the positivity Assumption S5. Therefore applying Hoeffding's inequality, for some positive ε: $J_{jj}^{(1)} > \varepsilon$.

We now focus on $J_{11}^{(1)} + J_{11}^{(2)} + J_{22}^{(2)}$. By algebra we have:

$$
\partial_{\beta_j}\mathcal{E}_A(t;\beta^*,S_c^*,\pi^*,Z) \quad = \quad -t\mathcal{E}_A(t;\beta^*,S_c^*,\pi^*,Z)\left\{1 - \mathcal{E}_A(t;\beta^*,S_c^*,\pi^*,Z)\right\}.
$$

We have:

$$\begin{aligned}
J_{11}^{(1)}+J_{11}^{(2)}+J_{22}^{(2)} &= +\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{A_i-\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)\}\,Y_i(t)A_i \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau t\,\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)\{1-\mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)\} \\
&\quad \times d\left\{\Lambda_1^*(t,Z_i;\beta_{j0})-\Lambda_{10}(t,Z_i)+\Lambda_2^*(t,Z_i;\beta_{j0})-\Lambda_{20}(t,Z_i)\right\}Y_i(t).
\end{aligned}$$

Similarly to before, if we look at the expected value, we have:

$$\begin{aligned}
&\mathrm{E}\left[J_{11}^{(1)}+J_{11}^{(2)}+J_{22}^{(2)}\right] \\
&= \int_0^\tau \mathrm{E}\left[\left\{A-\frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}A\mathrm{E}\{Y(t)|A,Z\}\right] \\
&- \int_0^\tau \mathrm{E}\left[\frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\left\{1-\frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}\right. \\
&\quad \left.\times\, td\left\{\Lambda_1^*(t,Z;\beta_{j0})-\Lambda_{10}(t,Z)+\Lambda_2^*(t,Z;\beta_{j0})-\Lambda_{20}(t,Z)\right\}E\{Y(t)|A,Z\}\right].
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathrm{E}\left[J_{11}^{(1)}+J_{11}^{(2)}+J_{22}^{(2)}\right] \\
&= \int_0^\tau \mathrm{E}\left(\frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\left\{1-\frac{\mathrm{E}\left[Ae^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}{\mathrm{E}\left[e^{-\sum_{j=1}^{2}\beta_{j0}At}S_{c0}(t|A,Z)|Z\right]}\right\}\right. \\
&\quad \left.\times\, e^{-\sum_{j=1}^{2}\Lambda_{j0}(t,Z)}\left[A-te^{-\sum_{j=1}^{2}\beta_{j0}At}S_c(t|A,Z)\sum_{j=1}^{2}d\left\{\Lambda_j^*(t,Z;\beta_{j0})-\Lambda_{j0}(t,Z)\right\}\right]\right).
\end{aligned}$$

Again, by Assumption S5 and S9, we can conclude that $\mathrm{E}\left[J_{11}^{(1)}+J_{11}^{(2)}+J_{22}^{(2)}\right]>\varepsilon$ and therefore, by Hoeffding's inequality that $J_{11}^{(1)}+J_{11}^{(2)}+J_{22}^{(2)}>\varepsilon+o_p(1)$.

- Case b): J is invertible if $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$.

We have:

$$|J^{(1)}| \quad = \quad (J_{11}^{(1)} + J_{11}^{(3)})(J_{22}^{(1)} + J_{22}^{(3)}),$$

We are now left to prove that $J_{jj}^{(1)} + J_{jj}^{(3)} \neq 0$ when $\Lambda^*(\cdot) = \Lambda_0(\cdot)$. We have:

$$J_{jj}^{(1)} + J_{jj}^{(3)} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{A_i - \mathcal{E}_{A_i}(t;\beta^*,S_c^*,\pi^*,Z_i)\} Y_i(t) \left\{A_i + \partial_{\beta_j} d\Lambda_j^*(t,Z_i)\right\},$$

and similarly to before, by Assumption S9 we can prove that $J_{jj}^{(1)} + J_{jj}^{(3)} > \varepsilon$. □

Proof of Theorem 7 (sketc). By Taylor expansion we have:

$$S_{2,n}(\beta,\hat{S}_c,\hat{\pi},\hat{\Lambda}) = S_{2,n}(\beta_0,\hat{S}_c,\hat{\pi},\hat{\Lambda}) + \nabla_\beta S_{2,n}(\beta^*,\hat{S}_c,\hat{\pi},\hat{\Lambda})\,(\beta-\beta_0)^\top, \tag{3.51}$$

where β^* lies between β and β_0.

Under Assumptions S1-6, it can be proved that:

$$\nabla_\beta S_{2,n}(\beta^*,\hat{S}_c,\hat{\pi},\hat{\Lambda}) = \nabla_\beta S_{2,n}(\beta^*,S_c^*,\pi^*,\Lambda^*) + o_p(1). \tag{3.52}$$

In the proof of Theorem 6 we moreover proved that:

$$\nabla_\beta S_{2,n}(\beta^*,S_c^*,\pi^*,\Lambda^*) \quad = \quad J + \begin{bmatrix} \beta_1^* - \beta_{10} \\ \beta_2^* - \beta_{20} \end{bmatrix} J',$$

where the above multiplication is intended componentwise and

$$J = \begin{bmatrix} J_{11}^{(1)} + J_{11}^{(2)} + J_{11}^{(3)} & J_{12}^{(1)} \\ J_{21}^{(1)} & J_{22}^{(1)} + J_{22}^{(2)} + J_{22}^{(3)} \end{bmatrix},$$

and

$$J' = \begin{bmatrix} J_{11}^{(1')} + J_{11}^{(2')} & J_{12}^{(1')} + J_{12}^{(2')} \\ J_{21}^{(1')} + J_{21}^{(2')} & J_{22}^{(1')} + J_{22}^{(2')} \end{bmatrix}.$$

We now focus on term $S_{2,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda})$. We have the following decomposition:

$$\begin{aligned} S_{2,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) &= +S_{2,n}(\beta_0, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) - S_{2,n}(\beta_0, S_c^*, \pi^*, \hat{\Lambda}) \\ &\quad +S_{2,n}(\beta_0, S_c^*, \pi^*, \hat{\Lambda}) - S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) \\ &\quad +S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) \\ &= Q_1 + Q_2 + Q_3. \end{aligned}$$

We remind the reader that in the previous part of the proof we proved that $\hat{\beta}^{(2)} - \beta_0 = O_p(n^{-\delta})$ for $|\delta| < 1/2$.

Putting all the above together, by definition of $\hat{\beta}$ we have:

$$\sqrt{n}(\hat{\beta}^{(2)} - \beta_0) = J^{-1}\sqrt{n}\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + Q_1 + Q_2\} + o_p(1). \tag{3.53}$$

$S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*)$ is by double robustness of the score, (Theorem 2), a sum of i.i.d. mean zero terms. Similarly to the proof of Theorem 4, we can prove that $Q_2 = o_p(n^{-1/2})$ if $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ and $Q_1 = o_p(n^{-1/2})$ if $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$. We therefore now divide the proof in three different cases according to which model is correctly specified.

- Case a): $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$, $\pi^*(\cdot) = \pi_0(\cdot)$ and $\Lambda^*(\cdot,\cdot) \neq \Lambda_0(\cdot,\cdot)$ with $a_n = n^{-1/2}$, $b_n = n^{-1/2}$.

As said before, we can prove that $Q_1 = o_p(n^{-1/2})$, therefore, by (3.53) and by Assumption A'1, we have:

$$\sqrt{n}(\hat{\beta} - \beta_0) \quad = \quad J^{-1}\sqrt{n}\left\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \frac{1}{n}\sum_{i=1}^{n}\sigma_{6i}\right\} + o_p(1).$$

$\sqrt{n}(\hat{\beta} - \beta_0)$ can be therefore written as sum of i.i.d mean zero terms, and therefore, by multivariate central limit theorem, it is asymptotically normal.

Part a) of the Theorem follows directly.

- Case b): $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$, $S_c^*(\cdot|\cdot,\cdot) \neq S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) \neq \pi_0(\cdot)$ with $c_n = n^{-1/2}$.

As said before, we can prove that $Q_2 = o_p(n^{-1/2})$, therefore, by (3.53) and by Assumption B' 1, we have:

$$\sqrt{n}(\hat{\beta} - \beta_0) \quad = \quad J^{-1}\sqrt{n}\left\{S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + \frac{1}{n}\sum_{i=1}^{n}\sigma_{7i}\right\} + o_p(1).$$

$\sqrt{n}(\hat{\beta} - \beta_0)$ can be therefore written as sum of i.i.d mean zero terms, and therefore, by multivariate central limit theorem, it is asymptotically normal.

Part b) of the Theorem follows directly.

- Case c): $S_c^*(\cdot|\cdot,\cdot) = S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot) = \pi_0(\cdot)$ and $\Lambda^*(\cdot,\cdot) = \Lambda_0(\cdot,\cdot)$ with $a_nc_n = o(n^{-1/2})$ and $b_nc_n = o(n^{-1/2})$.

In this case we have both $Q_1 = o_p(n^{-1/2})$ and $Q_2 = o_p(n^{-1/2})$. Therefore:

$$\sqrt{n}(\hat{\beta} - \beta_0) \quad = \quad J^{-1}\sqrt{n}S_{2,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) + o_p(1). \tag{3.54}$$

Moreover, when both models are correct, J simplifies to a diagonal matrix with diagonal element equals to $J_{jj}^{(1)} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{A_i - \mathcal{E}_{A_i}(t;\beta^*, S_c^*, \pi^*, Z_i)\} A_i Y_i(t)$.

MCLT can be applied to $\sqrt{n}S_{2,n}(\beta_0, S_{c0}, \pi_0, \Lambda_0)$ to prove asymptotic normality. Specifically we consider the following multivariate martingale $M_i(t) = [M_{1i}(t), M_{2i}(t)]^\top$ with respect to the filtration

$\mathcal{F}_t = \sigma\left\{N_{ji}(s), Y_i(s+), A_i, Z_i\ : j = 1,2,\ i = 1,\ldots,n,\ 0 < s < t\right\}$. We consider the following two-dimensional vector:

$$M^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t h(u; A_i, Z_i) dM_i(u),$$

where $h(t;A,Z) = A - \mathcal{E}_A(t;\beta, S_{c0}, \pi_0, Z)$. Since $h(t;A,Z)$ is predictable with respect to the filtration, then $M^n(t)$ is a multivariate martingale too. By Assumption S3, we have $< M_{1i}(t), M_{2i}(t) >=< M_{1i}(t), M_{1j}(t) >=< M_{2i}(t), M_{2j}(t) >=< M_{1i}(t), M_{2j}(t) >= 0$ for each $i \neq j$ therefore:

$$\begin{aligned} < M_1^n(t), M_2^n(t) > \quad &= \quad < \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t h(u; A_i, Z_i) dM_{1i}(u), \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t h(u; A_i, Z_i) dM_{2i}(u) > \\ &= \quad \frac{1}{n}\sum_{i,j=1}^n \int_0^t h^2(u; A_i, Z_i) d < M_{1i}(t), M_{2j}(t) >= 0, \end{aligned}$$

and so the two components of the multidimensional martingale $M^n(t)$ are orthogonal to each other. Therefore, we can apply the multidimensional version of the martingale central limit theorem of Rebolledo (Theorem 5 of Rebolledo (1978)).

First we verify Assumption 2 about the convergence of the variance. We have, by Assumption C'1, for $j = 1,2$:

$$\begin{aligned} < M_j^n(t), M_j^n(t) > &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t h^2(u;A_i,Z_i)d\Lambda_{j0}(u|A_i,Z_i)Y_i(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t h^2(u;A_i,Z_i)\left\{d\Lambda_{j0}(u,Z_i)+\beta_{j0}Adu\right\}Y_i(u) \\ &= \int_0^t \left\{P'(u)\beta_{j0}+Q'_j(u)\right\}du \xrightarrow{p} \int_0^t \left\{p'(u)\beta_{j0}+q'_j(u)\right\}du = V'_j(t), \end{aligned}$$

and so Assumption 2 of the MCLT is verified.

We now look at Assumption 1 about the jumps of each component of the martingale. Rebolledo (1978) at pag. 39 claims that if the Lindeberg condition is verified, then Assumption 1 of its Theorem holds. We therefore needs to prove that, for any ε and any j:

$$\int_0^\tau \frac{1}{n}\sum_{i=1}^{n} h^2(u;A_i,Z_i)\mathbb{1}\left\{|h(u;A_i,Z_i)| > \sqrt{n}\varepsilon\right\}Y_i(t)\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\} \xrightarrow{p} 0,$$

by Assumptions S1 and S5, we know that:

$$|h(t;A,Z)| \leq 1 + \frac{\max\{1, e^{-(\beta_{10}+\beta_{20})\tau}\}C_3}{\min\{1, e^{-(\beta_{10}+\beta_{20})\tau}\}C_1C_2+1-C_3} < \infty,$$

so, we have:

$$\begin{aligned} &\int_0^\tau \frac{1}{n}\sum_{i=1}^{n} h^2(u;A,Z)\mathbb{1}\left\{|h(u;A,Z)| > \sqrt{n}\varepsilon\right\}Y_i(t)\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\} \\ \leq\ &\int_0^\tau \frac{1}{n}\sum_{i=1}^{n} h^2(u;A,Z)\mathbb{1}\left\{1 + \frac{\max\{1, e^{-(\beta_{10}+\beta_{20})\tau}\}C_3}{\min\{1, e^{-(\beta_{10}+\beta_{20})\tau}\}C_1C_2+1-C_3} > \sqrt{n}\varepsilon\right\}Y_i(t) \\ &\times\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\}. \end{aligned}$$

Moreover, by Assumption S2, we also know that:

$$\begin{aligned}
&\left|\int_0^\tau \frac{1}{n}\sum_{i=1}^n h^2(u;A,Z)\mathbb{1}\left\{1+\frac{\max\{1,e^{-(\beta_{10}+\beta_{20})\tau}\}C_3}{\min\{1,e^{-(\beta_{10}+\beta_{20})\tau}\}C_1C_2+1-C_3}>\sqrt{n}\varepsilon\right\}\right.\\
&\left.\cdot Y_i(t)\left\{d\Lambda_{j0}(t,Z_i)+\beta_{j0}A_idt\right\}\right|\\
\leq\ &\left\{1+\frac{\max\{1,e^{-(\beta_{10}+\beta_{20})\tau}\}C_3}{\min\{1,e^{-(\beta_{10}+\beta_{20})\tau}\}C_1C_2+1-C_3}\right\}^2\\
&\times\mathbb{1}\left\{1+\frac{\max\{1,e^{-(\beta_{10}+\beta_{20})\tau}\}C_3}{\min\{1,e^{-(\beta_{10}+\beta_{20})\tau}\}C_1C_2+1-C_3}>\sqrt{n}\varepsilon\right\}\tau\left|L_j+\beta_{j0}\right|\xrightarrow{p}0,
\end{aligned}$$

and so Assumption 1 of the martingale central limit Theorem holds.

Therefore, we can conclude that

$$\sqrt{n}S_{2,n}(\beta_0,S_{c0},\pi_0,\Lambda_0)=\mathbb{M}^n(t)\xrightarrow{D}\mathcal{N}(0,V'(\tau)). \tag{3.55}$$

By the above and (3.54) part c) of the Theorem follows. □

Proof of Theorem 8 (sketch). This proof is similar to the proof of Theorem 5 and we leave it to the reader.

□

3.8.6 Proofs of Lemmas

Proof of Lemma 7. We remind the reader that:

$$S_{1,n}(\beta,S_c,\pi,\Lambda)=\left\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau e^{(\beta_1+\beta_2)A_it}S_c^{-1}(t|A_i,Z_i)\left\{A_i-\pi(Z_i)\right\}dM_{ji}(t;\beta_j,\Lambda_j)\right\}_{j=1,2}.$$

By algebra we have the following decomposition of the score:

$$\begin{aligned}
& S_{1,n}(\beta,\hat{S}_c,\hat{\pi},\hat{\Lambda}) \\
= \;& S_{2,n}(\beta,\hat{S}_c,\hat{\pi},\hat{\Lambda}) - S_{2,n}(\beta_0,\hat{S}_c,\hat{\pi},\hat{\Lambda}) \\
& + S_{2,n}(\beta_0,\hat{S}_c,\hat{\pi},\hat{\Lambda}) - S_{2,n}(\beta_0,\hat{S}_c,\hat{\pi},\Lambda^*) \\
& + S_{2,n}(\beta_0,\hat{S}_c,\hat{\pi},\Lambda^*) - S_{2,n}(\beta_0,S_c^*,\pi^*,\Lambda^*) \\
& + S_{2,n}(\beta_0,S_c^*,\pi^*,\Lambda^*) \\
= \;& Q^{(1)} + Q^{(2)} + Q^{(3)} + Q^{(4)}.
\end{aligned}$$

We first of all notice that by Assumption 1 and S7, we have:

$$\begin{aligned}
& \left[\hat{S}_c^{-1}(t|A_i,Z_i)\{A_i - \hat{\pi}(Z_i)\} - \{S_c^*(t|A_i,Z_i)\}^{-1}\{A_i - \pi^*(Z_i)\}\right] && (3.56) \\
\leq \; & \sup_{t\in[0,\tau],Z\in\mathcal{Z},A\in 0,1} \left|\hat{S}_c^{-1}(t|A,Z) - \{S_c^*(t|A,Z)\}^{-1}\right| && (3.57) \\
& + \sup_{t\in[0,\tau],Z\in\mathcal{Z},A\in 0,1} \left|\{S_c^*(t|A,Z)\}^{-1}\right| |\hat{\pi}(Z) - \pi^*(Z)| = o_p(1). && (3.58)
\end{aligned}$$

Moreover, we notice that, by Assumption 1 and S7 $K^{(1)}(\beta_0,\pi^*,S_c^*) = O_p(1)$. By Assumptions S1, S7, S6, we have: $K_j^{(2)}(\beta,\pi^*,S_c^*) = O_p(|\beta_j - \beta_{j0}|)$.

We now work on each term separately.

- Term $Q^{(1)}$:

Algebra and the application of Lemma 12 gives us:

$$
\begin{aligned}
Q_j^{(1)} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{e^{(\beta_1+\beta_2)A_it} - e^{(\beta_{10}+\beta_{20})A_it}\right\}\hat{S}_c^{-1}(t|A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\} \\
&\quad\times\left\{dN_{ji}(t)-Y_i(t)\beta_jA_idt-Y_i(t)d\hat{\Lambda}_j(t,Z_i;\beta)\right\} \\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_it}\hat{S}_c^{-1}(t|A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\} \\
&\quad\times\left\{dN_{ji}(t)-Y_i(t)\beta_jA_idt-Y_i(t)d\hat{\Lambda}_j(t,Z_i;\beta)\right\} \\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_it}\hat{S}_c^{-1}(t|A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\} \\
&\quad\times\left\{dN_{ji}(t)-Y_i(t)\beta_{j0}A_idt-Y_i(t)d\hat{\Lambda}_j(t,Z_i;\beta_0)\right\}.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
Q_j^{(1)} &= (\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_1^*+\beta_2^*)A_it}A_itK_{ji}^{(3)}(t,\beta,\hat{S}_c\hat{\pi}) \\
&\quad-(\beta_j-\beta_{j0})K^{(1)}(\beta_0,\hat{S}_c,\hat{\pi})-K_j^{(2)}(\beta,\hat{S}_c,\hat{\pi}),
\end{aligned}
$$

for some β^* between β_0 and β.

Moreover:

$$
\begin{aligned}
Q_j^{(1)} &= (\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_1^*+\beta_2^*)A_it}A_itK_{ji}^{(3)}(t,\beta,S_c^*,\pi^*) \qquad (3.59)\\
&\quad +(\beta_1+\beta_2-\beta_{10}-\beta_{20})^2\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_1^{**}+\beta_2^{**})A_it}A_it^2\left\{K_{ji}^{(3)}(t,\beta,\hat{S}_c,\hat{\pi})-K_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)\right\}\\
&\quad +(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_it}A_it\left\{K_{ji}^{(3)}(t,\beta,\hat{S}_c,\hat{\pi})-K_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)\right\}\\
&\quad -(\beta_j-\beta_{j0})K^{(1)}(\beta_0,S_c^*,\pi^*)\\
&\quad -(\beta_j-\beta_{j0})\left\{K^{(1)}(\beta,\hat{S}_c,\hat{\pi})-K^{(1)}(\beta_0,S_c^*,\pi^*)\right\}\\
&\quad -K_j^{(2)}(\beta,S_c^*,\pi^*)\\
&\quad +K_j^{(2)}(\beta,\hat{S}_c,\hat{\pi})-K_j^{(2)}(\beta,S_c^*,\pi^*),
\end{aligned}
$$

where β_j^{**} is a point between β^* and β_0. We remind the reader that the quantities $K^{(1)},K_j^{(2)},K_j^{(3)},K_j^{(4)}$ are defined in equations (3.34)-(3.37).

We work now on term $\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_1^*+\beta_2^*)A_it}A_itK_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)$. By algebra and by Lemma 12:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_1^*+\beta_2^*)A_it}A_itK_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)\\
= \;&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_{10}+\beta_{20})A_it}A_it\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}dM_{ji}(t;\beta,\hat{\Lambda})\\
&+(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau e^{(\beta_1^{**}+\beta_2^{**})A_it}A_it^2\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}dM_{ji}(t;\beta,\hat{\Lambda}).
\end{aligned}
$$

Therefore:

$$\begin{aligned}
= \; & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)} dM_{ji}(t) \\
& -K_j^{(4)}(\beta_0,\pi^*,S_c^*,\Lambda^*)-(\beta_j-\beta_{j0})K^{(1)}(\beta_0,S_c^*,\pi^*) \\
& -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)} d\left\{\hat{\Lambda}_j(t,Z_i;\beta)-\hat{\Lambda}_j(t,Z_i;\beta_0)\right\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)} d\left\{\hat{\Lambda}_j(t,Z_i;\beta_0)-\Lambda_j^*(t,Z_i)\right\} \\
& +(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_1^{**}+\beta_2^{**})A_i t} A_i t^2 \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)} dM_{ji}(t;\beta,\hat{\Lambda}).
\end{aligned}$$

We work now on term $\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left\{K_{ji}^{(3)}(t,\beta,\hat{S}_c,\hat{\pi})-K_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)\right\}$. By algebra and by Lemma 12:

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left\{K_{ji}^{(3)}(t,\beta,\hat{S}_c,\hat{\pi})-K_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)\right\} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left[\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)}-\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right] dM_{ji}(t;\beta,\hat{\Lambda})
\end{aligned}$$

Therefore

$$\begin{aligned}
= \quad & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left[\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)} - \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right] dM_{ji}(t;\beta_0,\Lambda^*) \\
& -(\beta_j-\beta_{j0})\left\{K^{(1)}(\beta_0,\hat{S}_c,\hat{\pi}) - K^{(1)}(\beta_0,S_c^*,\pi^*)\right\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left\{\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)} - \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right\} Y_i(t) \\
& \times d\left\{\hat{\Lambda}_j(t,Z_i;\beta) - \Lambda_j^*(t,Z_i;\beta_0)\right\} \\
& -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left[\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)} - \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right] Y_i(t) \\
& \times d\left\{\hat{\Lambda}_j(t,Z_i;\beta_0) - \Lambda_j^*(t,Z_i)\right\}.
\end{aligned} \tag{3.60}$$

Therefore putting together (3.59), (3.60) and (3.60), we get:

$$\begin{aligned}
& Q_j^{(1)} \\
& = Q_j^{(11)} + Q_j^{(12)} + Q_j^{(13)} + Q_j^{(14)} + Q_j^{(15)} + Q_j^{(16)} + Q_j^{(17)} + Q_j^{(18)} + Q_j^{(19)} + Q_j^{(110)} + Q_j^{(111)} + Q_j^{(112)},
\end{aligned}$$

where

$$Q_j^{(11)} = (\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t} A_i t \left\{\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)} - \frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right\} dM_{ji}(t),$$

$$\begin{aligned}
Q_j^{(12)} = & -(\beta_1+\beta_2-\beta_{10}-\beta_{20})K_j^{(4)}(\beta_0,S_c^*,\pi^*,\Lambda^*) \\
& -(\beta_1+\beta_2-\beta_{10}-\beta_{20})(\beta_j-\beta_{j0})K^{(1)}(\beta_0,S_c^*,\pi^*),
\end{aligned}$$

$$Q_j^{(13)} = -(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}A_i t\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}$$

$$\times\, d\left\{\hat{\Lambda}_j(t,Z_i;\beta_j)-\Lambda_j^*(t,Z_i;\beta_{j0})\right\},$$

$$Q_j^{(14)} = -(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}A_i t\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}$$

$$\times\, d\left\{\hat{\Lambda}_j(t,Z_i;\beta_{j0})-\Lambda_j^*(t,Z_i)\right\},$$

$$Q_j^{(15)} = (\beta_1+\beta_2-\beta_{10}-\beta_{20})^2\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_1^{**}+\beta_2^{**})A_i t}A_i t^2\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}dM_{ji}(t;\beta,\hat{\Lambda}),$$

$$Q_j^{(16)} = +\frac{1}{n}(\beta_1+\beta_2-\beta_{10}-\beta_{20})^2\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_1^{**}+\beta_2^{**})A_i t}A_i t^2$$

$$\times\left\{K_{ji}^{(3)}(t,\beta,\hat{S}_c.\hat{\pi})-K_{ji}^{(3)}(t,\beta,S_c^*,\pi^*)\right\},$$

$$Q_j^{(17)} = +(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}A_i t\left\{\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)}-\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right\}$$

$$\times\, dM_{ji}(t;\beta_0,\Lambda^*),$$

$$Q_j^{(18)} = -(\beta_j-\beta_{j0})(\beta_1+\beta_2-\beta_{10}-\beta_{20})\left\{K^{(1)}(\beta_0,\hat{S}_c,\hat{\pi})-K^{(1)}(\beta_0,S_c^*,\pi^*)\right\},$$

$$
\begin{aligned}
Q_j^{(19)} = & -(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}A_i t\left\{\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)}-\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right\} \\
& \times Y_i(t)d\left\{\hat{\Lambda}_j(t,Z_i;\beta)-\Lambda_j^*(t,Z_i;\beta_0)\right\},
\end{aligned}
$$

$$
\begin{aligned}
Q_j^{(110)} = & -(\beta_1+\beta_2-\beta_{10}-\beta_{20})\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} e^{(\beta_{10}+\beta_{20})A_i t}A_i t\left\{\frac{A_i-\hat{\pi}(Z_i)}{\hat{S}_c(t|A_i,Z_i)}-\frac{A_i-\pi^*(Z_i)}{S_c^*(t|A_i,Z_i)}\right\} \\
& \times Y_i(t)d\left\{\hat{\Lambda}_j(t,Z_i;\beta_0)-\Lambda_j^*(t,Z_i)\right\},
\end{aligned}
$$

$$
Q_j^{(111)} = -(\beta_j-\beta_{j0})K^{(1)}(\beta_0,S_c^*,\pi^*)-(\beta_j-\beta_{j0})\left\{K^{(1)}(\beta,\hat{S}_c,\hat{\pi})-K^{(1)}(\beta_0,S_c^*,\pi^*)\right\},
$$

$$
Q_j^{(112)} = -K_j^{(2)}(\beta,S_c^*,\pi^*)+K_j^{(2)}(\beta,\hat{S}_c,\hat{\pi})-K_j^{(2)}(\beta,S_c^*,\pi^*).
$$

$Q_j^{(11)}$ is a martingale integral, therefore, by Lemma 13, we have $Q_j^{(11)} = O_p(n^{-1/2}|\beta_1+\beta_2-\beta_{10}-\beta_{20}|)$.

By Assumptions S1 and S7, we have $Q_j^{(12)} = O_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\,|\beta_j-\beta_{j0}|\right)$ and $Q_j^{(16)} = O_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|^2\right)$.

By Assumptions S1, S7 and S6 we have $Q_j^{(13)} = O_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\,|\beta_j-\beta_{j0}|\right)$.

By Assumptions 1, S1, S7 and (3.56) we have $Q_j^{(14)} = o_p(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|)$, $Q_j^{(17)} = o_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|^2\right)$ and $Q_j^{(18)} = o_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\,|\beta_j-\beta_{j0}|\right)$.

By Assumptions 1 and (3.56) we have $Q_j^{(17)} = o_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\right)$. Moreover, we notice that, if $\Lambda^*(t,Z)=\Lambda^0(t,Z)$, we would have $Q_j^{(17)} = o_p\left(n^{-1/2}|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\right)$ since $dM(t;\beta_0,\Lambda^*)$ would be a martingale.

By Assumption 1 and S6, we have $Q_j^{(19)} = o_p\left(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\left|\beta_j-\beta_{j0}\right|\right)$. Moreover by Cauchy Schwartz inequality, together with Assumption S6 we get $Q_j^{(110)} = o_p\left(n^{-1/2}|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\left|\beta_j-\beta_{j0}\right|\right)$.

By Assumption S1, we have $Q_j^{(111)} = -(\beta_j-\beta_{j0})K^{(1)}(\beta_0,S_c^*,\pi^*)+o_p\left(\left|\beta_j-\beta_{j0}\right|\right)$.

Moreover, by Assumptions S6 and S(3.56), we have $Q_j^{(112)} = -K_j^{(2)}(\beta,S_c^*,\pi^*)+o_p\left(\left|\beta_j-\beta_{j0}\right|\right)$. We moreover notice that, if $S_c^*(\cdot|\cdot,\cdot)=S_{0c}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot)=\pi_0(\cdot)$, by Lemma 14, we have $Q_j^{(112)} = -K_j^{(2)}(\beta,S_c^*,\pi^*)+O_p\left(n^{-1/2}|\beta-\beta_0|\right)$.

Therefore:

$$\begin{aligned}
Q_j^{(1)} &= -(\beta_1+\beta_2-\beta_{10}-\beta_{20})K_j^{(4)}(\beta_0,S_c^*,\pi^*,\Lambda^*)\\
&\quad -(\beta_j-\beta_{j0})K_j^{(1)}(\beta_0,S_c^*,\pi^*)-K_j^{(2)}(\beta,S_c^*,\pi^*)\\
&\quad +O_p\left(n^{-1/2}|\beta_1+\beta_2-\beta_{10}-\beta_{20}|+|\beta_1+\beta_2-\beta_{10}-\beta_{20}|^2\right)\\
&\quad +o_p(|\beta_1+\beta_2-\beta_{10}-\beta_{20}|+n^{-1/2}|\beta_1+\beta_2-\beta_{10}-\beta_{20}|\left|\beta_j-\beta_{j0}\right|).
\end{aligned}$$

- Term $Q^{(2)}$:

Adding and subtracting we have:

$$\begin{aligned}
Q_j^{(2)} &= \frac{1}{n}\sum_{i=1}^n\int_0^\tau e^{(\beta_{10}+\beta_{20})A_it}\left\{S_c^*(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi^*(Z_i)\right\}Y_i(t)d\left\{\hat\Lambda_j(t,Z_i;\beta_0)-\Lambda_j^*(t,Z_i)\right\}\\
&\quad +\frac{1}{n}\sum_{i=1}^n\int_0^\tau e^{(\beta_{10}+\beta_{20})A_it}\left[\hat S_c^{-1}(t|A_i,Z_i)\left\{A_i-\hat\pi(Z_i)\right\}-\left\{S_c^*(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi^*(Z_i)\right\}\right]\\
&\quad \times Y_i(t)d\left\{\hat\Lambda_j(t,Z_i;\beta_0)-\Lambda_j^*(t,Z_i)\right\}\\
&= Q_j^{(21)}+Q_j^{(22)}.
\end{aligned}$$

By Assumption 1, S1, S7, we have:

$$\begin{aligned}\left|Q_j^{(21)}\right| &\leq \left|\int_0^\tau \left[\frac{1}{n}\sum_{i=1}^n e^{(\beta_{10}+\beta_{20})A_i t}\left\{S_c^*(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi^*(Z_i)\right\}Y_i(t)\right]\right.\\ &\qquad \left.\sup_{Z\in\mathcal{Z}} d\left\{\hat{\Lambda}_j(t,Z;\beta_0)-\Lambda_j^*(t,Z)\right\}\right|\\ &= o_p(1).\end{aligned}$$

We moreover notice that, if $S_c^*(\cdot|\cdot,\cdot)=S_{c0}(\cdot|\cdot,\cdot)$ and $\pi^*(\cdot)=\pi_0(\cdot)$, by Lemma 14, we have $Q_j^{(21)}=o_p(n^{-1/2})$. Otherwise, by Lemma 10, we have $Q_j^{(21)}=O_p(n^{-1/2})$ under Assumptions B1 and B2.

Moreover by Cauchy-Schwartz inequality we have:

$$\begin{aligned}&\left|Q_j^{(22)}\right|\\ \leq\ & \frac{1}{n}e^{(\beta_{10}+\beta_{20})\tau}\sqrt{\sum_{i=1}^n \sup_{t\in[0,\tau]}\left[\hat{S}_c^{-1}(t|A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\}-\left\{S_c^*(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi^*(Z_i)\right\}\right]^2}\\ &\cdot\sqrt{\sum_{i=1}^n\left[\int_0^\tau d\left\{\hat{\Lambda}_j(t,Z_i;\beta_0)-\Lambda_j^*(t,Z_i)\right\}\right]^2}.\end{aligned}$$

Therefore:

$$\begin{aligned}&\left|Q_j^{(22)}\right|\\ \leq\ & e^{(\beta_{10}+\beta_{20})\tau}\sqrt{\frac{1}{n}\sum_{i=1}^n \sup_{t\in[0,\tau]}\left[\hat{S}_c^{-1}(t|A_i,Z_i)\left\{A_i-\hat{\pi}(Z_i)\right\}-\left\{S_c^*(t|A_i,Z_i)\right\}^{-1}\left\{A_i-\pi^*(Z_i)\right\}\right]^2}\\ &\cdot\left\{\sup_{t\in[0,\tau],Z\in\mathcal{Z}}\left|\hat{\Lambda}_j(\tau,Z;\beta_0)-\Lambda_j^*(\tau,Z)\right|\right\}.\end{aligned}$$

Therefore, by Assumption 1 and by the fact that $a_n c_n = o_p(n^{-1/2})$ and $b_n c_n = o_p(n^{-1/2})$ we have $Q_j^{(22)} = o_p(n^{-1/2})$. Therefore $Q_j^{(2)} = o_p(1) + o_p(n^{-1/2})$.

- Term $Q^{(3)}$:

By Assumption 1 we have $Q^{(3)} = o_p(1)$. Moreover, we notice that, if $\Lambda^*(\cdot,\cdot) = \Lambda^0(\cdot,\cdot)$, we have $Q^{(3)} = o_p(n^{-1/2})$ since it would be a martingale integral with integrand converging to zero. Otherwise, $Q^{(3)} = O_p(n^{-1/2})$ under Assumptions A1, A2.

Putting all of these steps together we have:

$$\begin{aligned} S_{1,n}(\beta, \hat{S}_c, \hat{\pi}, \hat{\Lambda}) &= S_{1,n}(\beta_0, S_c^*, \pi^*, \Lambda^*) \\ &\quad -(\beta_1+\beta_2-\beta_{10}-\beta_{20})K^{(4)}(\beta_0, S_c^*, \pi^*, \Lambda^*) \\ &\quad -(\beta_j-\beta_{j0})K^{(1)}(\beta_0, S_c^*, \pi^*) - K^{(2)}(\beta, S_c^*, \pi^*) \\ &\quad +O_p\left(n^{-1/2}|\beta_1+\beta_2-\beta_{10}-\beta_{20}| + |\beta_1+\beta_2-\beta_{10}-\beta_{20}|^2\right) + Q^{(21)} + Q^{(3)}, \end{aligned}$$

where $Q^{(21)} = o_p(n^{-1/2}), Q^{(3)} = o_p(n^{-1/2})$ under case c) of the Theorem, $Q^{(21)} = o_p(n^{-1/2}), Q^{(3)} = O_p(n^{-1/2})$ under case a) of the Theorem, and under case c), $Q^{(21)} = O_p(n^{-1/2}), Q^{(3)} = o_p(n^{-1/2})$.

If $\hat{\Lambda}(\cdot,\cdot)$ depends on the unknown β, by Assumption S6, we have:

$$\begin{aligned} K_j^{(2)}(\beta, S_c^*, \pi^*) &= (\beta_j - \beta_{j0})\frac{1}{n}\sum_{i=1}^n \int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t} \{S_c^*(t|A_i, Z_i)\}^{-1} \{A_i - \pi^*(Z_i)\} Y_i(t) \\ &\quad \times \frac{1}{n}\sum_{l=1}^n q_{jl}(t)dt, \end{aligned}$$

where we call $q_{ji}(t)$ a function, such that:

$$\hat{\Lambda}_j(t, Z; \beta) - \hat{\Lambda}_j(t, Z; \beta_0) = (\beta_j - \beta_{j0}) * \frac{1}{n}\sum_{i=1}^n \int_0^\tau q_{ji}(t). \qquad (3.61)$$

Therefore, we can conclude that:

$$\begin{aligned} S_{1,n}(\beta,\hat{S}_c,\hat{\pi},\hat{\Lambda}) &= S_{1,n}(\beta_0,S_c^*,\pi^*,\Lambda^*)+Q^{(21)}+Q^{(3)}+K(\beta-\beta_0) \\ &+O_p\left(n^{-1/2}\left|\beta_1+\beta_2-\beta_{10}-\beta_{20}\right|+\left|\beta_1+\beta_2-\beta_{10}-\beta_{20}\right|^2\right), \end{aligned}$$

where K is a 2X2 matrix with the following components:

$$K_{jj} = -K^{(1)}(\beta_0,S_c^*,\pi^*)-K_j^{(2)}(\beta_0,S_c^*,\pi^*)/(\beta_j-\beta_{j0})-K_j^{(4)}(\beta_0,S_c^*,\pi^*,\Lambda^*),$$

and

$$K_{12} = -K_4^{(1)}(\beta_0,S_c^*,\pi^*,\Lambda^*), \quad K_{21} = -K_j^{(4)}(\beta_0,S_c^*,\pi^*,\Lambda^*).$$

Remark 3. *If the estimator $\hat{\Lambda}(\cdot,\cdot)$ does not depend on β or it depends on some initial estimator of it the decomposition simplifies. Specifically, terms $K_j^{(2)}(\beta,S_c^*,\pi^*),Q^{(12)},Q^{(18)},Q^{(112)},Q^{(113)}$ cancels.*

□

Proof of Lemma 8. We remind the reader that
$Q_j^{(3)}=\sqrt{n}\left[S_j^n(\beta_0,\hat{S}_c,\hat{\pi},\Lambda^*)-\sqrt{n}S_j^n(\beta_0,S_{c0},\pi_0,\Lambda^*)\right]$. Using the fact that $\hat{\pi}(z)=\text{expit}(\hat{\alpha}^\top z)$, $\hat{S}_c(t|a,z)=\exp\left(-\hat{\Lambda}_c(t)e^{\hat{\eta}^\top d}\right)$ we have, by Taylor expansion:

$$\begin{aligned} Q_j^{(3)} &= \sqrt{n}\left[S_j^n(\beta_0,\hat{\eta},\hat{\Lambda}_c,\hat{\alpha},\Lambda^*)-\sqrt{n}S_j^n(\beta_0,\eta_0,\Lambda_{c0},\alpha_0,\Lambda^*)\right] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t}\left\{D_{ij}^n(t,\eta_0,\Lambda_{c0},\alpha_0\right\}^\top \Delta dM_{ji}(t;\beta_{j0},\Lambda_j^*)+o_p(1), \end{aligned}$$

where

$$D^n_{ij}(t,\eta,\Lambda_c,\alpha) := [\partial_\eta f(t,A_i,Z_i;\eta,\Lambda_c,\alpha), \partial_{\Lambda_c} f(t,A_i,Z_i;\eta,\Lambda_c,\alpha), \partial_\alpha f(t,A_i,Z_i;\eta,\Lambda_c,\alpha)]^\top,$$

$$f(t,A_i,Z_i;\eta,\Lambda_c,\alpha) := \left\{A_i - \text{expit}(\alpha^\top Z)\right\} \exp\left(\Lambda_c(t) e^{\eta^\top D}\right),$$

and $\Delta := [\hat{\eta} - \eta_0, \hat{\Lambda}_c(t) - \Lambda_{c0}(t), \hat{\alpha} - \alpha_0]^\top$.

Standard algebra gives us:

$$\begin{aligned} D^n_{ij}(t,\eta,\Lambda_c,\alpha) &= [\left\{A_i - \text{expit}(\alpha^\top Z_i)\right\}) \exp\left(\Lambda_c(t) e^{\eta^\top D_i}\right) \Lambda_c(t) e^{\eta^\top D_i} D_i, \\ &\quad \left\{A_i - \text{expit}(\alpha^\top Z_i)\right\}) \exp\left(\Lambda_c(t) e^{\eta^\top D_i}\right) e^{\eta^\top D_i}, \\ &\quad , -\exp\left(\Lambda_c(t) e^{\eta^\top D_i}\right) \text{expit}(\alpha^\top Z_i) e^{\alpha^\top Z_i} Z_i]^\top. \end{aligned}$$

Moreover, we know by traditional theory that

$$\hat{\alpha} - \alpha_0 = O_p(n^{-1/2}), \quad \hat{\eta} - \eta_0 = O_p(n^{-1/2}), \quad \sup_{t\in[0,\tau]} \left\{\hat{\Lambda}_c(t) - \Lambda_{c0}(t)\right\} = O_p(n^{-1/2}).$$

Therefore, by the above and by Assumption A*1 we have:

$$\begin{aligned} Q^{(3)}_j &= \sqrt{n} \int_0^\tau \left[\left\{P_1^{(a')}\right\}^\top (t)(\hat{\eta} - \eta_0) + P_2^{(a')}(t)(\hat{\Lambda}_c(t;\hat{\eta}) - \Lambda_{c0}(t)) - \left\{P_3^{(a')}\right\}^\top (t)(\hat{\alpha} - \alpha_0)\right] dt \\ &= \sqrt{n} \int_0^\tau \left[\left\{p_1^{(a')}\right\}^\top (t)(\hat{\eta} - \eta_0) + p_2^{(a')}(t)(\hat{\Lambda}_c(t;\hat{\eta}) - \Lambda_{c0}(t)) - \left\{p_3^{(a')}\right\}^\top (t)(\hat{\alpha} - \alpha_0)\right] dt \\ &\quad + o_p(1). \end{aligned} \tag{3.62}$$

Lemma 15 and 16 provide the influence functions of $\hat{\alpha}, \hat{\eta}, \hat{\Lambda}_c$. Therefore, plugging them in

(3.62) we can conclude that:

$$
\begin{aligned}
\sqrt{n}Q_j^{(3)} \quad = \quad & \left(\int_0^\tau \left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)} - \left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right] s_d^{(0)}(t) d\Lambda_{c0}(t)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\} dM_i^c(t) \\
& \times \left[\int_0^\tau (p_1^{(a')})^\top(t) - [\int_0^\tau p_2^{(a')}(t) \int_0^t d\Lambda_{c0}(u;\eta_0)\frac{s_d^{(1)}(u)}{s_d^{(0)}(u)} dt\right] \\
& + \int_0^\tau p_2^{(a')}(t)\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \left\{s_d^{(0)}(u)\right\}^{-1} dM_i^c(u) \\
& - \int_0^\tau (p_3^{(a')})^\top(t)\left(E\left[Z^\top Z\pi_0(Z_i)\left\{1-\pi_0(Z_i)\right\}\right]\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \left\{A_i - \pi_0(Z_i)\right\} dt + o_p(1).
\end{aligned}
$$

□

Proof of Lemma 9. We have:

$$
\begin{aligned}
\sqrt{n}Q_j^{(21)} \quad = \quad & -\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau e^{(\beta_{10}+\beta_{20})A_i t} \left\{S_c^*(t|A_i,Z_i)\right\}^{-1} \left\{A_i - \pi^*(Z_i)\right\} Y_i(t) \\
& \times \left[(\hat{\gamma}_j - \gamma_{j0})^\top Z_i dt + d\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\}\right].
\end{aligned}
$$

We notice that, by Lin and Ying (1994b), under regularity Assumptions, we have, for each t, z:

$$
\left\{\hat{\Lambda}_j(t,z;\beta_{j0},\hat{\gamma}_j) - \Lambda_{j0}(t,z)\right\} = O_p(n^{-1/2}), \tag{3.63}
$$

and

$$
\left\{\hat{\gamma}_j - \gamma_{j0}\right\} = O_p(n^{-1/2}). \tag{3.64}
$$

Therefore, by the above and by Assumption B*1, we have:

$$Q_j^{(21)} = \sqrt{n}\int_0^\tau \left[(\hat{\gamma}_j - \gamma_{j0})^\top p_1^{(b')}(t)dt + p_0^{(b')}(t)d\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\}\right] + o_p(1). \tag{3.65}$$

Lemma 17 and 18 provide influence functions for $\hat{\gamma}_j$ and $\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j)$.

Therefore, plugging them into (3.65) we have;

$$\begin{aligned} Q_j^{(21)} &= \int_0^\tau \left(\left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{ D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)} \right\}^{\otimes 2} Y_i(t)dt \right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{ Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)} \right\} dM_{ji}(t) \right)^\top \\ &\quad \cdot \left\{ p_1^{(b')}(t)dt - p_0^{(b')}(t)\frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}dt \right\} \\ &\quad + \int_0^\tau p_0^{(b')}(t)\left\{s_z^{(0)}(t)\right\}^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n dM_{ji}(t) + o_p(1). \end{aligned}$$

□

Proof of Lemma 10. The proof is similar to the proof of Lemma S9, using Lemma 19 instead of S18 and we leave it to the reader. □

3.8.7 Additional Lemmas and proofs

Lemma 11. *Let's consider a generic probability model $p(x;\beta_0,\eta_0)$ for which β_0 is the true parameter of interest and η_0 is the nuisance parameter. Let $\phi(x,\beta,\eta)$ be such that $E_{\beta,\eta}\{\phi(x,\beta,\eta)\} = 0$ and let $\Lambda^\perp$ be the space orthogonal to the nuisance tangent space. Then, $\phi \in \Lambda^\perp$ if and only if the score is orthogonal, that is*

$$\frac{\partial}{\partial r}E\{\phi(x,\beta_0,\eta^r)\}|_{r=0} = 0, \tag{3.66}$$

where $\eta^r = \eta_0 + r\Delta\eta$.

Proof of Lemma 11. We have:

$$\int \phi(x, \beta_0, \eta^r) p(x; \beta_0, \eta^r) dx = 0,$$

and so

$$\begin{aligned} 0 &= \left. \frac{\partial}{\partial r} \int \phi(x, \beta_0, \eta^r) p(x; \beta_0, \eta^r) dx \right|_{r=0} \\ &= \int \partial_r \phi(x, \beta_0, \eta^r)|_{r=0}\, p(x; \beta_0, \eta^r)|_{r=0}\, dx + \int \phi(x, \beta, \eta^r)|_{r=0}\, \partial_r p(x; \beta_0, \eta^r)|_{r=0}\, dx \\ &= \int \partial_r \phi(x, \beta_0, \eta^r)|_{r=0}\, p(x; \beta_0, \eta_0) dx + \int \phi(x, \beta_0, \eta_0) \partial_r \log p(x; \beta_0, \eta^r)|_{r=0}\, p(x; \beta_0, \eta_0) dx \\ &= \frac{\partial}{\partial r} \mathrm{E}\left\{\phi(x, \beta_0, \eta^r)\right\}|_{r=0} + \mathrm{E}\left\{\phi(x, \beta_0, \eta_0) S_\eta\right\}. \end{aligned}$$

Therefore, if $\phi \in \Lambda^\perp$, and therefore $\mathrm{E}\left\{\phi(x, \beta_0, \eta_0) S_\eta\right\} = 0$, we obtain $\frac{\partial}{\partial r}\mathrm{E}\left\{\phi(x, \beta_0, \eta^r)\right\}|_{r=0} = 0$. On the other hand, if $\frac{\partial}{\partial r}\mathrm{E}\left\{\phi(x, \beta_0, \eta^r)\right\}|_{r=0} = 0$, we have $\mathrm{E}\left\{\phi(x, \beta_0, \eta_0) S_\eta\right\} = 0$ and so $\phi \in \Lambda^\perp$. □

Lemma 12. *A simple application of the multidimensional mean value theorem gives us*

$$e^{(\beta_1+\beta_2)t} - e^{(\beta_{10}+\beta_{20})t} = e^{(\beta_1^*+\beta_2^*)t} t\left(\beta_1 + \beta_2 - \beta_{10} - \beta_{20}\right),$$

where β_j^* *is a point between* β_j *and* β_{j0} *for* $j = 1, 2$.

Lemma 13. *Let* $H_i(t)$ *be a random variable such that* $P\left(\sup_{i=1,\ldots,n; t\in[0,\tau]} |H_i(t)| \le K\right) = 1$ *for some* $K < \infty$. *We have, for any bounded* β_j*:*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau H_i(t) dM_{ji}(t; \beta_j, \Lambda_j) = E\left[\int_0^\tau H(t) dM_j(t; \beta_j, \Lambda_j)\right] + O_p(n^{-1/2}) \qquad (3.67)$$

Proof of Lemma 13. By definition of dM_{ji}, we have:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} H_i(t)dM_{ji}(t;\beta_j,\Lambda_j) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} H_i(t)\left[dN_{ji}(t)-Y_i(t)\left\{\beta_j A_i dt + d\Lambda_j(t,Z_i)\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta_i H_i(X_i) - X_i\beta_j A_i - \int_0^{X_i} H_i(t)d\Lambda_j(t,Z_i).\end{aligned}$$

We have, by Assumptions S1 and S2

$$\left|\delta_i H_i(X_i) - X_i\beta_j A_i - \int_0^{X_i} H_i(t)d\Lambda_j(t,Z_i)\right| \leq K + \tau|\beta_j| + K|\Lambda_j(\tau,Z_i)| < \infty$$

Therefore, by Hoeffding's inequality we have (3.67). □

Lemma 14. *It holds:*

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{A_i-\pi_0(Z_i)\right\}\left\{S_{c0}(t|A_i,Z_i)\right\}^{-1}Y_i(t)e^{(\beta_{10}+\beta_{20})A_i t}\right| = O_p\left(n^{-1/2}\right).$$

Proof of Lemma 14. This is a slightly modified version of Lemma A13 of Hou et al. (2021), adapted to include the survival of the censoring. We leave the proof to the reader. □

Lemma 15. *Let $\pi(Z;\alpha) = expit(\alpha^{\top}Z)$ and let $\hat{\alpha}$ be the MLE estimator for α. We have:*

$$\sqrt{n}(\hat{\alpha}-\alpha_0) = \left(E\left[Z^{\top}Z\pi_0(Z)\left\{1-\pi_0(Z)\right\}\right]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\left\{A_i-\pi_0(Z_i)\right\}+o_p(1).$$

Proof of Lemma 15. Estimation of parameter α is done through classical MLE method. By classical

MLE argument we have (proved in Zeng and Chen (2010)):

$$
\begin{aligned}
\sqrt{n}(\hat{\alpha}-\alpha_0) &= \left\{-\frac{1}{n}\sum_{i=1}^{n}\frac{e^{-(\alpha_0)^\top Z_i}}{\left\{1+e^{-(\alpha_0)^\top Z_i}\right\}^2}Z_i^\top Z_i\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{A_i e^{-(\alpha_0)^\top Z_i}-1+A_i}{1+e^{-(\alpha_0)^\top Z_i}}Z_i \\
&= \left(\mathrm{E}\left[Z^\top Z\pi_0(Z)\{1-\pi_0(Z)\}\right]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\{A_i-\pi_0(Z_i)\}+o_p(1).
\end{aligned}
$$

□

Lemma 16. *Let* $S_c(t|A,Z)=g(t|A,Z;\eta,\Lambda_c)=\exp\left(-\Lambda_c e^{\eta^\top D}\right)$ *and let* $\hat{\eta}$ *and* $\hat{\Lambda}_c(t)$ *be the Cox estimators. Under Assumptions S2, S4 and A*3 we have:*

$$
\begin{aligned}
&\sqrt{n}\{\hat{\eta}-\eta_0\} \\
= &\left(\int_0^\tau\left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)}-\left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right]s_d^{(0)}(t)d\Lambda_{c0}(t)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}dM_i^c(t)+o_p(1),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sqrt{n}\left\{\hat{\Lambda}_c(t;\hat{\eta})-\Lambda_{c0}(t)\right\} \qquad (3.68)\\
= &\left\{\left(\int_0^\tau\left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)}-\left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right]s_d^{(0)}(t)d\Lambda_{c0}(t)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}dM_i^c(t)\right\}^\top \\
&\cdot\int_0^t -d\Lambda_{c0}(u;\eta_0)\frac{s_d^{(1)}(u)}{s_d^{(0)}(u)}du+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t\left\{s_d^{(0)}(u)\right\}^{-1}dM_i^c(u)+o_p(1).
\end{aligned}
$$

Proof of Lemma 16. Estimation of parameter η uses the following score:

$$\begin{aligned}
U_1(\eta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i - \frac{\sum_{j=1}^{n} Y_j(t) D_j e^{\eta^\top D_j}}{\sum_{j=1}^{n} Y_j(t) e^{\eta^\top D_j}}\right\} dN_i^c(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i - \frac{\sum_{j=1}^{n} Y_j(t) D_j e^{\eta^\top D_j}}{\sum_{j=1}^{n} Y_j(t) e^{\eta^\top D_j}}\right\} \left[dM_i^c(t) + Y_i(t) d\Lambda_{c0}(t) e^{(\eta_0)^\top D_i} dt\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i - \frac{\sum_{j=1}^{n} Y_j(t) D_j e^{\eta^\top D_j}}{\sum_{j=1}^{n} Y_j(t) e^{\eta^\top D_j}}\right\} dM_i^c(t),
\end{aligned}$$

where $M_i^c(t) = N_i^c(t) - Y_i(t)\Lambda_{c0}(t) e^{(\eta_0)^\top D_i}$, and $N^c(t) := \mathbf{1}\{X \leq t, \delta = 0\}$. By Taylor expansion we have:

$$\begin{aligned}
U_1(\eta_0) &= U_1(\eta_0) - U_1(\hat{\eta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\hat{\eta} - \eta_0\}^\top \left[\frac{S_d^{(2)}(t)}{S_d^{(0)}(t)} - \left\{\frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}^2\right] dM_i^c(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\hat{\eta} - \eta_0\}^\top \left[\frac{S_d^{(2)}(t)}{S_d^{(0)}(t)} - \left\{\frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}^2\right] dN_i^c(t).
\end{aligned}$$

Therefore, by Assumption A*2

$$
\begin{aligned}
&\sqrt{n}\{\hat{\eta}-\eta_0\} \\
= \;& \left(\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\frac{S_d^{(2)}(t)}{S_d^{(0)}(t)}-\left\{\frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}^2\right]dN_i^c(t)\right)^{-1} \\
&\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i-\frac{\sum_{j=1}^{n}Y_j(t)D_je^{\eta^\top D_j}}{\sum_{j=1}^{n}Y_j(t)e^{\eta^\top D_j}}\right\}dM_i^c(t) \\
= \;& \left(\frac{1}{n}\sum_{i=1}^{n}\int_0^{t}\left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)}-\left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right]dN_i^c(t)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}dM_i^c(t)+o_p(1) \\
= \;& \left(\int_0^{\tau}\left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)}-\left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right]s_d^{(0)}(t)d\Lambda_{c0}(t)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}dM_i^c(t) \\
&+o_p(1)..
\end{aligned}
\tag{3.69}
$$

We now need to find the influence function of $\hat{\Lambda}_c(t;\hat{\eta})-\Lambda_{c0}(t)=\hat{\Lambda}_c(t;\hat{\eta})-\hat{\Lambda}_c(t;\eta_0)+\hat{\Lambda}_c(t;\eta_0)-\Lambda_{c0}(t)$. Since $\hat{\Lambda}_c(t;\eta)=\int_0^t\frac{\sum_{i=1}^{n}dN_i^c(u)}{\sum_{i=1}^{n}Y_i(u)e^{\eta^\top D_i}}$, by Taylor expansion and by (3.69) and Assumption A* 3 we have:

$$
\begin{aligned}
&\sqrt{n}\left\{\hat{\Lambda}_c(t;\hat{\eta})-\hat{\Lambda}_c(t;\eta_0)\right\} \\
&=-(\hat{\eta}-\eta_0)^\top\int_0^t d\hat{\Lambda}_c(u;\eta_0)\frac{S_d^{(1)}(u)}{S_d^{(0)}(u)}du \\
&=\left\{\left(\int_0^{\tau}\left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)}-\left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right]s_d^{(0)}(t)d\Lambda_{c0}(t)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau}\left\{D_i-\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}dM_i^c(t)\right\}^\top \\
&\quad\times\int_0^t -d\hat{\Lambda}_c(u;\eta_0)\frac{s_d^{(1)}(u)}{s_d^{(0)}(u)}du+o_p(1).
\end{aligned}
\tag{3.70}
$$

Estimation of parameter $\Lambda_c(t)$ uses the following score:

$$\begin{aligned} U_2(\Lambda_c(t);\eta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{dN_i^c(u) - Y_i(t)d\Lambda_c(u)e^{\eta^\top D_i}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left[dM_i^c(u) - Y_i(u)d\left\{\Lambda_c(u)e^{\eta^\top D_i} - \Lambda_{c0}(u)e^{(\eta_0)^\top D_i}\right\}\right]. \end{aligned}$$

Therefore, by construction of $\hat{\Lambda}_c(t;\eta_0)$ we have

$$U_2(\Lambda_{c0}(t);\eta_0) = U_2(\Lambda_{c0}(t);\eta_0) - U_2(\hat{\Lambda}_{c0}(t;\eta_0);\eta_0) = \int_0^t S_d^{(0)}(t)\left\{d\hat{\Lambda}_{c0}(t;\eta_0) - d\Lambda_{c0}(t)\right\},$$

and so we have:

$$\hat{\Lambda}_c(t;\eta_0) - \Lambda_{c0}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{s_d^{(0)}(u)\right\}^{-1} dM_i^c(u) + o_p(1),$$

Therefore, by putting together (3.70) and (3.71) we get:

$$\begin{aligned} &\sqrt{n}\left\{\hat{\Lambda}_c(t;\hat{\eta}) - \Lambda_{c0}(t)\right\} \qquad (3.71) \\ = &\left\{\left(\int_0^\tau \left[\frac{s_d^{(2)}(t)}{s_d^{(0)}(t)} - \left\{\frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^2\right] s_d^{(0)}(t)d\Lambda_{c0}(t)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\} dM_i^c(t)\right\}^\top \\ &\times \int_0^t -d\Lambda_{c0}(u;\eta_0)\frac{s_d^{(1)}(u)}{s_d^{(0)}(u)}du + \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{s_d^{(0)}(u)\right\}^{-1} dM_i^c(u) + o_p(1). \end{aligned}$$

□

Lemma 17. *Let* $\Lambda_j(t,Z) = G_j(t) + \gamma_j^\top Zt$ *and let* γ_j *be estimated using* (3.10) *of the paper. Under*

*Assumption B*2 it holds:*

$$\begin{aligned}
&\sqrt{n}(\hat{\gamma}_j - \gamma_{j0})^\top \\
= \quad &\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2} Y_i(t)dt\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \left\{Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}\right\} dM_{ji}(t) + o_p(1).
\end{aligned}$$

Proof of Lemma 17. The parameter γ_j is estimated through the following score:

$$\begin{aligned}
&U_1\left([\beta_j^{in}, \gamma_j]^\top\right) \\
= \quad &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}\left\{dN_{ji}(t) - Y_i(t)\beta_j^{in} A_i dt - Y_i(t)\gamma_j^\top Z_i dt\right\} \\
= \quad &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\} \\
&\times\left\{dM_{ji}(t) + Y_i(t)dG_{j0}(t) - Y_i(t)(\beta_j^{in} - \beta_{j0})A_i dt - Y_i(t)(\gamma_j - \gamma_{j0})^\top Z_i dt\right\} \\
= \quad &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}\left\{dM_{ji}(t) - Y_i(t)(\beta_j^{in} - \beta_{j0})A_i dt - Y_i(t)(\gamma_j - \gamma_{j0})^\top Z_i dt\right\}.
\end{aligned}$$

Here β_j^{in} is just some initial β_j that we need for technical reason.

Therefore, by construction, we have:

$$\begin{aligned}
U_1\left([\beta_{j0}, \gamma_{j0}]^\top\right) &= U_1\left([\beta_j^{in}, \hat{\gamma}_j]^\top\right) - U_1\left([\beta_{j0}, \gamma_{j0}]^\top\right) && (3.72) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}\left\{Y_i(t)[\beta_j^{in} - \beta_{j0}, \hat{\gamma}_j - \gamma_{j0}]^\top D_i dt\right\}. && (3.73)
\end{aligned}$$

Therefore, by Assumption B*2, we have:

$$
\begin{aligned}
& \sqrt{n}(\hat{\gamma}_j - \gamma_{j0})^\top \qquad (3.74)\\
= & \left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{D_i - \frac{S_d^{(1)}(t)}{S_d^{(0)}(t)}\right\}^{\otimes 2} Y_i(t)dt\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{Z_i - \frac{S_z^{(1)}(t)}{S_z^{(0)}(t)}\right\} dM_{ji}(t)\\
= & \left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2} Y_i(t)dt\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}\right\} dM_{ji}(t) + o_p(1).
\end{aligned}
$$

(3.75)

□

Lemma 18. *Let* $\Lambda_j(t,Z) = G_j(t) + \gamma_j^\top Zt$ *and let* γ_j *be estimated using* (3.10) *of the paper and* $G_j(t)$ *be estimated using* (3.11). *Under Assumptions B*2 it holds:*

$$
\begin{aligned}
& \sqrt{n}\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\}\\
= & \int_0^t \left\{s_z^{(0)}(u)\right\}^{-1} \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n dM_{ji}(u) - \left(\left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2} Y_i(t)dt\right]^{-1}\right.\right.\\
& \left.\left. \times \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}\right\} dM_{ji}(t)\right)^\top s_z^{(1)}(u)du\right].
\end{aligned}
$$

Proof of Lemma 18. The nuisance parameter $G_j(t)$ is estimated through the following score:

$$
\begin{aligned}
U_2(G_j(t);\beta_{j0},\hat{\gamma}_j) &= \frac{1}{n}\sum_{i=1}^n \int_0^t \left\{dN_i(t) - Y_i(t)dG_j(t) - Y_i(t)\beta_{j0}A_i dt - Y_i(t)\hat{\gamma}_j Z_i dt\right\}\\
&= \frac{1}{n}\sum_{i=1}^n \int_0^t \left[dM_{ji}(t) - Y_i(t)d\left\{G_j(t) - G_{j0}(t)\right\} - Y_i(t)(\hat{\gamma}_j - \gamma_{j0})^\top Z_i dt\right].
\end{aligned}
$$

Therefore by construction we have:

$$\begin{aligned} U_2(G_{j0}(t);\beta_{j0},\hat{\gamma}_j) &= U_2(G_{j0}(t);\beta_{j0},\gamma_{j0}) - U_2(\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j);\beta_{j0},\hat{\gamma}_j) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t Y_i(t) d\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\}, \end{aligned}$$

and therefore, by Assumption B*2 and Lemma 17:

$$\begin{aligned} \sqrt{n}\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\} &= \sqrt{n}\int_0^t \left\{S_z^{(0)}\right\}^{-1} \frac{1}{n}\sum_{i=1}^{n}\left[dM_{ji}(t) - Y_i(t)(\hat{\gamma}_j - \gamma_{j0})^\top Z_i dt\right] \quad (3.76) \\ &= \int_0^t \left\{s_z^{(0)}(t)\right\}^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} dM_{ji}(t) - \sqrt{n}(\hat{\gamma}_j - \gamma_{j0})^\top s_z^{(1)}(t) dt\right]. \end{aligned}$$

Hence, by (3.74), we have:

$$\begin{aligned} &\sqrt{n}\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\} \quad (3.77) \\ = \; &\sqrt{n}\int_0^t \left\{S_z^{(0)}(u)\right\}^{-1} \frac{1}{n}\sum_{i=1}^{n}\left[dM_{ji}(u) - Y_i(u)(\hat{\gamma}_j - \gamma_{j0})^\top Z_i du\right] \\ = \; &\int_0^t \left\{s_z^{(0)}(u)\right\}^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} dM_{ji}(u)\right. \\ &-\left(\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2} Y_i(t)dt\right]^{-1}\right. \quad (3.78) \\ &\left.\left.\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \left\{Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}\right\} dM_{ji}(t)\right)^\top s_z^{(1)}(u)du\right]. \end{aligned}$$

□

Lemma 19. *Let* $\Lambda_j(t,Z) = G_j(t) + \gamma_j^\top Zt$ *and let* γ_j *be estimated using* (3.10) *of the paper and* $G_j(t)$

be estimated using (3.12). *Under Assumption B*3 it holds:*

$$
\begin{aligned}
&\sqrt{n}\left\{\hat{G}_j(t;\beta_{j0},\hat{\gamma}_j) - G_{j0}(t)\right\} \\
= \quad & \int_0^t \left\{s_{wz}^{(0)}(u;S_c^*,\pi^*)\right\}^{-1} \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n dM_{ji}(u) - \left(\left[\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{D_i - \frac{s_d^{(1)}(t)}{s_d^{(0)}(t)}\right\}^{\otimes 2} Y_i(t)dt\right]^{-1}\right.\right. \\
& \left.\left.\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{Z_i - \frac{s_z^{(1)}(t)}{s_z^{(0)}(t)}\right\} dM_{ji}(t)\right)^\top s_{wz}^{(1)}(u;S_c^*,\pi^*)du\right].
\end{aligned}
$$

Proof of Lemma 19. The proof is similar to the proof of Lemma 18 and we leave it to the reader. □

3.9 Acknowledgements

This research was supported by a grant from the National Institute on Aging; NIH/NIA R03 AG062432. Chapter 3, is currently being prepared for submission for publication of the material. Rava, Denise; Xu, Ronghui. The dissertation author was the primary investigator and author of this material.

Chapter 4

Doubly Robust Estimation under the Marginal Structural Cox Model for a Binary Treatment

4.1 Introduction

4.1.1 Background

In the analysis of time-to-event data it is often of interest to estimate the causal treatment effect. For a binary treatment, this compares the failure time that would be observed if a patient were treated with the failure time that would be observed if a patient were untreated or received a different treatment. These hypothetical failure times are called potential outcomes. The causal treatment effect is usually summarized by the causal hazard ratio, i.e. the ratio between the hazards of the two potential outcomes. In clinical trials, the causal hazard ratio is used to compare the survival of a patient if s/he had been given a specific treatment with the survival of the same patient if s/he had been given another treatment or a placebo. The causal hazard ratio can also be employed in therapeutic trials to asses if a treatment can shorten the duration of the illness.

Here we aim at estimating the causal hazard ratio, assumed constant under the Marginal Structural Cox model (Hernán et al., 2001), as a concise summary of the effect of a treatment on a survival endpoint. The Marginal Structural Cox model has been widely used in observational studies for the analysis of the effect of different therapies on the progress of various diseases, such as AIDS, hemodialysis and HIV (Cole et al., 2003; Feldman et al., 2004; Sterne et al., 2005; Hernán et al., 2006; Buchanan et al., 2014).

Marginal Structural models (MSMs), introduced by Robins et al. (2000), model the marginal distribution of the potential outcome of interest. The term structural specifies that the model is posed on the potential outcome of interest and the term marginal refers to the fact that MSMs do not incorporate confounders. These characteristics warrant the causal interpretation of MSMs' coefficients. This remains true even for noncollapsible models, i.e. models that change their forms and their parameters when a covariate is integrated out. Martinussen and Vansteelandt (2013) study the non-collapsibility of the Cox model, reason why the causal hazard ratio cannot be estimated simply incorporating the confounders into the model. Algebra shows indeed that the causal hazard ratio does not equal the expected value of the ratio between the hazards of the treated and the untreated given the confounders.

In randomized trials, the absence of confounders of the relationship between treatment and outcome guarantees consistent estimation of MSMs' parameters using standard regression methods. However, randomized trials can be infeasible in practice and often the assumption of absence of confounders is not realistic. Under the assumption of no unmeasured confounders, inverse probability weighting (IPW) has been widely used to perform estimation of the causal parameters characterizing this type of models (Robins et al., 2000; Hubbard et al., 2000; Hernán et al., 2001; Chen and Tsiatis, 2001; Lunceford and Davidian, 2004; Wei, 2008; Zhang and Schaubel, 2011; Buchanan et al., 2014). IPW adjusts for the confounders weighting every observation via the inverse of its propensity score, that is the probability of receiving the treatment conditional to the confounders. Propensity scores are unknown in observational studies, therefore they need to be estimated. A major drawback of this method is that, mistakes in the estimation of the propensity score, induces bias in the estimation of the causal effect of interest. To overcome this drawback, augmented inverse probability weighting (AIPW) has been introduced. AIPW estimators have the interesting property of being doubly robust, that is of being consistent as long as one of two models is correctly specified (Robins and Rotnitzky, 2001; Wang and Chen, 2001; Van Der Laan et al., 2003; Bang and Robins, 2005; Tsiatis, 2006; Tchetgen Tchetgen et al., 2010; Zhang and Schaubel, 2012a; Jiang et al., 2017; Hou et al., 2021; Tan, 2019). Usually these estimators are doubly robust with respect to the treatment assignment and the conditional outcome model.

In this paper, we derive the AIPW estimator for the Marginal Structural Cox Model that is doubly robust with respect to the propensity score and the survival function conditional to the treatment and the confounders. To this aim, we augment the Cox-IPW estimators of both structural parameters, hazard ratio and baseline hazard, offering protection against possible misspecification of the propensity score. To the best of our knowledge, this is the first time a doubly robust estimator for this model is proposed.

Our estimator is model-doubly robust, that is, it needs only one of the two models to be correctly specified to be consistent. Moreover it is asymptotically normal when only one of the two models is correctly specified and estimated at the classical $\sqrt{n}$ rate of convergence. The proposed estimator is also rate-doubly robust, that is, when both models are correctly specified, it only needs their product rates to be $\sqrt{n}$ to be consistent and asymptotically normal. Therefore, our estimator, does not require the propensity score or the conditional outcome model to be estimated at a specific rate and in theory, one of the two rates can be extremely slow as long as the other one makes up for it. This characteristic allows the user to use, for the estimation of the propensity score and/or the conditional outcome model, nonparametric methods as boosted logistic regression, SVM, random survival forest and spline, that are known to converge with a rate slower than $\sqrt{n}$. This is particularly interesting since these methods relax the modeling assumptions typical of parametric and semiparametric methods giving one the possibility of overcoming the non-collapsibility challenge posed by the Marginal Structural Cox Model.

In simulations we will show how our estimator outperforms the existing IPW-Cox estimator both in terms of consistency and efficiency for different combinations of parametric and nonparametric estimators for the propensity score and the conditional outcome model.

4.1.2 Related work

The literature on AIPW and doubly robust estimators for both structural and non structural quantities of interest is rapidly growing. We name here few significative examples of both groups.

Robins (1998) derive a generic class of semiparametric estimators for the parameters of MSMs with a focus on efficient estimators. For the marginal structural Cox model they propose an augmented version of the Cox-IPW estimator of the hazard ratio. However, they don't augment the IPW estimators of both structural parameters, hazard ratio and baseline hazard. As a consequence, their estimator is not robust against possible misspecification of the propensity score. Zhang and Schaubel (2012b) and Bai et al. (2017) propose AIPW estimators separately for $E[f\{T(1)\}]$ and $E[f\{T(0)\}]$ for different f. Yang et al. (2020) derive a doubly robust estimator for the structural failure time model following the theory of Bickel et al. (1993). Sjölander and Vansteelandt (2017) develop a doubly robust estimator for the attributable fraction $1-\frac{1-S_{T(0)}(t)}{1-S(t)}$. Even thought Zhang and Schaubel (2012b), Bai et al. (2017) and Yang et al. (2020) estimators are model-doubly robust, they are not rate-doubly robust; their asymptotic normality relies on the $\sqrt{n}$ convergence of classical semiparametric estimators of the nuisance parameters.

Cui et al. (2020) adapt causal forest to the survival framework of censored data for estimation of heterogeneous treatment effect using AIPWC methodology. Their estimator is however not doubly robust. Dukes et al. (2019b) and Hou et al. (2021) propose doubly robust estimators for the hazard difference in low and high-dimension, respectively. Both of these works exploit the good property of the hazard difference of being collapsible and they focus on the estimation of the conditional hazard difference. Bickel and Kwon (2001); Tchetgen Tchetgen et al. (2010); Tan (2019) claim that no DR estimating function exists for the conditional logistic regression model on the observed data with respect to the outcome model and the propensity score. However, Tchetgen Tchetgen et al. (2010); Tan (2019), exploiting a different parametrization, propose a doubly robust estimator with respect to different models.

4.1.3 Organization of the paper

In Section 2 we define the notation, the model and the assumptions we work with. In Section 3 we derive the AIPW estimator while in Section 4 we explain its asymptotic properties. In Section 5 we study the finite sample properties of our estimator through extensive simulations. In Section 6 we apply our estimator to the data from a cohort of Japanese men in Hawaii followed since the 1960s in order to study the effect of mid-life alcohol exposure on overall death. We conclude with the discussion in the last section. We report the proofs of all the results in Section 4.8.2 of the Supplementary Material.

4.2 Marginal Structural Cox Model

Let A be a binary non randomized treatment and let $T(0), T(1)$ be the potential failure time of a subject if s/he had been untreated or treated, respectively. As usual we indicate with $\lambda(t)$ the hazard function. We assume that $T(a)$ follows a Marginal Structural Cox model, that is, for $a = 0, 1$:

$$\lambda_{T(a)}(t) = \lambda_0(t)\exp(\beta a), \tag{4.1}$$

where $\lambda_0(t)$ is an unknown function of t and β is the parameter of interest. We aim at estimating the treatment effect described by the log ratio of the hazards function of the two potential outcomes $T(1)$ and $T(0)$. We therefore focus on estimating the constant:

$$\beta = \log\left[\lambda_{T(1)}(t)\left\{\lambda_{T(0}(t)\right\}^{-1}\right]. \tag{4.2}$$

As is typical for time-to-event outcomes, the potential failure times $T(1), T(0)$ are subject to right censoring $C(1), C(0)$. As usual, we use $X(a) = \min\{T(a), C(a)\}$ to indicate the potential censored failure times, $\delta(a) = \mathbf{1}\{T(a) \leq C(a)\}$ the potential event indicators and Z the observed baseline covariates.

Ideally, we would be able to observe each subject under both treatment 1 and 0 and we would then have as full data $\{X(1), X(0), \delta(1), \delta(0), Z\}$. In practice, the two potential outcomes $\{X(0), \delta(0)\}$ and $\{X(1), \delta(1)\}$ are never observed simultaneously; indeed, if a subject is treated, only $\{X(1), \delta(1), A = 1, Z\}$ is observed and if a subject is not treated, only $\{X(0), \delta(0), A = 0, Z\}$ is observed. We use T, C, X, δ to indicate the corresponding observed failure, censoring, censored time and event indicator, respectively and we assume the following:

Assumption 2 (Consistency)**.** $T = T(a)$, $C = C(a)$, $X = X(a)$, $\delta = \delta(a)$ *if* $A = a$.

Assumption 3 (SUTVA)**.** *The potential outcomes on one unit are not affected by the treatment assignment of the other units.*

The consistency and the stable unit treatment value assumptions are typical of the causal inference literature (Robins et al., 2000; Hernán et al., 2001).

Because full data are not available in practice, we cannot consistently estimate the parameter of interest β using standard regression methods. Moreover, because the treatment is not randomized, we cannot assume that the group of treated subjects is a random sample of the population and therefore, $\lambda_{T(1)}(t)$ cannot simply be estimated using only the available treated subjects; the same can be said for $\lambda_{T(0)}(t)$. However, estimation of the parameter of interest from the observed data is possible under the following assumptions:

Assumption 4 (No unmeasured confounders)**.** $P(A = 1|X(1), X(0), \delta(1), \delta(0), Z) = P(A = 1|Z)$.

Assumption 5 (Positivity)**.** *There exists* $\varepsilon > 0$*, such that, for each z:* $\varepsilon < P(A = 1|Z = z) < 1 - \varepsilon$.

Assumption 4, also known as missing at random, assumes that the treatment assignment mechanism only depends on the observed covariates Z (Robins et al., 2000; Hernán et al., 2001). Assumption 5 assumes that every unit in the population has a chance of receiving each treatment (Rosenbaum and Rubin, 1983).

We allow the censoring to depend on the treatment because in reality the treatment might affect the censoring rate. For example, the side effects of a treatment might increase the percentage of treated patients that drop out of the study. Viceversa, the beneficial effect of a treatment might lower the percentage of treated subjects that are lost to follow-up. We assume the potential censoring times to obey the following:

Assumption 6 (Independent Censoring)**.** $C(a) \perp (T(a), Z)$ *for* $a = 0, 1$.

Above the symbol $\perp$ indicates independence. Classical survival analysis models pose assumptions on the hazard of the failure time conditional to some covariates. In these cases it is typical to assume $C(a) \perp T(a)|Z$; i.e. independence between the censoring and the failure time given the covariates incorporated in the model (Andersen and Gill, 1982; Bai et al., 2017). Since our model is on the marginal distribution of $T(a)$, we assume $C(a) \perp T(a)$. We moreover require $C(a) \perp Z$. In the literature of doubly robust estimators it is not uncommon to require assumptions on the censoring that are slightly stronger than the one needed for classical estimators (Dukes et al., 2019b; Hou et al., 2021). In practice, our assumption, that only requires the potential censoring to be independent of the covariates, is not unrealistic. This could be relaxed by using inverse probability of censoring weighting (Scharfstein and Robins, 2002), imposing a model on the censoring mechanism. We further analyze this in the discussion.

4.3 Augmented IPW score

We derive the AIPW score for β following the theory of Van Der Laan et al. (2003) and Tsiatis (2007). We start by constructing a full data estimating function, i.e. the estimating function we would use if we were to have observed for each individual i, $(X(1), X(0), \delta(1), \delta(0), A, Z)$. Model (4.1) is the intersection of the following two models: $\lambda_{T(0)}(t) = \lambda_0(t)$ and $\lambda_{T(1)}(t) = \lambda_0(t)\exp(\beta)$. We define:

$$\begin{aligned} M^0(t) &= N^0(t) - Y^0(t)\Lambda_0(u), && (4.3) \\ M^1(t) &= N^1(t) - Y^1(t)\Lambda_0(u)\exp(\beta), && (4.4) \end{aligned}$$

where $N^a(t) = \mathbf{1}\{X(a) \leq t, \delta(a) = 1\}$ and $Y^a(t) = \mathbf{1}\{X(a) \geq t\}$ for $a = 0, 1$. The quantities $M^0(t)$ and $M^1(t)$ are martingales with respect to the filtration $\mathcal{F}_t{}^a = \{N^a(u), Y_i^a(u^+) : \quad i = 1, \ldots, n,$

$0 < u < t\}$.

We can use as full data estimating function for β and $\Lambda_0(t)$, $U(h,t) = \sum_{a=0,1}\int_0^\tau h(a,t,u)dM^a(u)$ for a two-dimensional function h. We choose $h(a,t,u) = [a, \mathbf{1}\{u \le t\}]^\top$ obtaining as full data estimating function $U^F = [U_1^F, U_2^F(\cdot)]^\top$, where:

$$U_1^F = \sum_{a=0,1}\int_0^\tau a dM^a(t) \quad U_2^F(t) = \sum_{a=0,1}\int_0^t dM^a(u). \tag{4.5}$$

We now use U^F as a starting point to define, as observed data estimating function, $U^{IPW} = [U_1^{IPW}, U_2^{IPW}(\cdot)]^\top$ where:

$$U_1^{IPW} = \int_0^\tau wA dM(t) \quad U_2^{IPW}(t) = \int_0^t w dM(u), \quad M(t) = AM^1(t) + (1-A)M^0(t), \tag{4.6}$$

where $w = A/P(A=1|Z) + (1-A)/P(A=0|Z)$ is the usual inverse probability weight. The above technique, known as IPW, weights every treated and untreated observations by the inverse of the conditional probability of being treated and untreated, respectively. The weighted observations create a pseudopopulation where the treatment is randomized and in which $\lambda_T(t|A=a)$ is the same as $\lambda_{T(a)}(t)$ of the true population. We notice that by consistency we have $M(t) = N(t) - Y(t)\Lambda_0(t)\exp(\beta A)$ where $N(t) = \mathbf{1}\{X \le t, \delta = 1\}$ and $Y(t) = \mathbf{1}\{X \ge t\}$. We remark that while $M^1(t)$ and $M^0(t)$ are martingales, $M(t)$ is not a martingale since the assumed model (4.1) is on the potential outcomes and not on the observed one. We notice that the above U^{IPW} is the usual Cox-IPW score.

The estimating function U^{IPW} has been proved to be unbiased when the weights w are correctly estimated and therefore when the propensity score $P(A=1|Z)$ is known or estimated correctly (Hernán et al., 2001). However, when it is not correct, U^{IPW} is biased.

To protect against possible misspecification of the propensity score, we now augment U^{IPW} to obtain a doubly robust estimating function U^{AIPW}. Following the theory of Van Der Laan et al. (2003) we consider:

$$U^{AIPW} = U^{IPW} - \Pi\left\{U^{IPW}|\mathcal{T}\right\}. \tag{4.7}$$

Here $\mathcal{T}$ is the propensity score tangent space, the space spanned by the score of the propensity score and we indicate with $\Pi\{q(\cdot)|\mathcal{T}\}$ the projection of a function $q(\cdot)$ onto the space $\mathcal{T}$ in the Hilbert space with covariance inner product. In the following lemma we derive $\Pi\left\{U^{IPW}|\mathcal{T}\right\}$.

Lemma 20. *Under Assumption 4, for each $t \in [0,\tau]$:*

$$\begin{aligned}
\Pi\left\{U_1^{IPW}|\mathcal{T}\right\} &= \int_0^\tau \left[wAdE\{M(t)|A,Z\} - dE\{M(t)|A=1,Z\}\right], \qquad (4.8)\\
\Pi\left\{U_2^{IPW}(t)|\mathcal{T}\right\} &= \int_0^t \left[wdE\{M(u)|A,Z\} - dE\{M(u)|A=1,Z\} - dE\{M(u)|A=0,Z\}\right].
\end{aligned}$$

Applying the above lemma we derive the following AIPW estimating function $U^{AIPW} = [U_1^{AIPW}, U_2^{AIPW}(\cdot)]^\top$ where:

$$\begin{aligned}
U_1^{AIPW} &= \int_0^\tau wAdM(t) - wAd\mathrm{E}\{M(t)|A,Z\} + d\mathrm{E}\{M(t)|A=1,Z\}, \qquad (4.9)\\
U_2^{AIPW}(t) &= \int_0^t wdM(u) - wd\mathrm{E}\{M(u)|A,Z\} + d\mathrm{E}\{M(u)|A=1,Z\} + d\mathrm{E}\{M(u)|A=0,Z\}. \qquad (4.10)
\end{aligned}$$

We notice that we augment both components of the estimating function for estimation of both β and $\Lambda_0(t)$. Even thought β is the parameter of interest, $\Lambda_0(t)$ is still a structural parameter and so, to estimate $\Lambda_0(t)$ from the observed data, one needs to properly adjust for confounders. To protect against possible misspecification of the propensity score, it is therefore necessary to use an augmented estimator also for $\Lambda_0(t)$.

We now focus on the quantity $\mathrm{E}\{M(t)|A,Z\}$. Assumptions 2, 4 and 6 implies $T \perp C|A,Z$. Therefore we have $\mathrm{E}\{Y(t)|A,Z\} = S(t|A,Z)G(t|A)$ and $\mathrm{E}\{N(t)|A,Z\} = \int_0^t G(u|A)d\{1-S(u|A,Z)\}$ where $S(u|A,Z)$ and $G(u|A)$ are the conditional survivorship functions of T and C, respectively. Hence we get:

$$\begin{aligned} \mathrm{E}\{M(t)|A,Z\} &= \int_0^t G(u|A)d\{1-S(u|A,Z)\} \\ &\quad -\int_0^t d\Lambda_0(u)\exp(\beta A)S(u|A,Z)G(u|A). \end{aligned} \tag{4.11}$$

From now on we will use the propensity score notation $\pi(Z) = P(A=1|Z)$ and the shorthand $\pi_i = \pi(Z_i)$. Moreover we use the notation $U^{AIPW} = U^{AIPW}(\beta,\Lambda_0;\pi,S,G)$ to stress the dependency of the score on the nuisance parameters π,S,G. We will use β^o,π^o,S^o,G^o to indicate the true quantities. Score U^{AIPW}, for estimation of β and Λ_0, is doubly robust with respect to S and π.

Theorem 9. *Under Assumptions 2-6:* $E\left[U^{AIPW}(\beta^o,\Lambda_0^o;\pi,S,G^o)\right] = 0$ *if either* $S = S^o$ *or* $\pi = \pi^o$.

If we have n observations (X_i,δ_i,A_i,Z_i), we solve

$$\frac{1}{n}\sum_{i=1}^{n}\left\{U^{AIPW}\right\}_i = 0 \tag{4.12}$$

to estimate β and Λ_0. Here we use the notation $\left\{U^{AIPW}\right\}_i$ to indicate the two-dimensional score U^{AIPW} evaluated at X_i,δ_i,A_i,Z_i for observation i. We define:

$$\begin{aligned} S_i(t,a) = S(t|A=a \quad , \quad Z_i),\ G(t,a) = G(t|A=a), \\ R_i(t,S,G) \quad = \quad Y_i(t) - S_i(t,A_i)G(t,A_i), \end{aligned} \tag{4.13}$$

and for $l = 0, 1$,

$$\mathcal{S}^{(l)}(t; \beta, \pi, S, G) = \frac{1}{n} \sum_{i=1}^{n} \left[w_i \exp(\beta A_i) A_i^l R_i(t, S, G) + \sum_{a=0,1} a^l \exp(\beta a) S_i(t, a) G(t, a) \right]. \tag{4.14}$$

Solving for $\Lambda_0(t)$ we obtain, for each $t \in [0, \tau]$:

$$\tilde{\Lambda}_0(t; \beta, \pi, S, G) = \int_0^t \frac{\sum_{i=1}^n \left[w_i \left\{ dN_i(u) + G(u, A_i) dS_i(u, A_i) \right\} - \sum_{a=0,1} G(u, a) dS_i(u, a) \right]}{n \cdot \mathcal{S}^{(0)}(u; \beta, \pi, S, G)},$$

where we use the notation $\tilde{\Lambda}_0$ when β, π, S, G are assumed known and fixed. The above estimator is an augmented version of the IPW-Breslow estimator: $\int_0^t \sum_{i=1}^n w_i dN_i(u) \left\{ \sum_{i=1}^n w_i Y_i(u) \exp(\beta A_i) \right\}^{-1}$. The IPW-Breslow estimator is a consistent estimator for $\Lambda_0(t)$ when the propensity score is correctly estimated. The proposed augmented version protects against possible misspecification of the propensity score.

Finally, profiling out $\Lambda_0(t)$ we obtain the following AIPW score for estimation of β:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau w_i \left\{ A_i - \bar{A}(t; \beta, \pi, S, G) \right\} \left\{ dN_i(t) + G(t, A_i) dS_i(t, A_i) \right\} \\ - \int_0^\tau \sum_{a=0,1} \left\{ a - \bar{A}(t; \beta, \pi, S, G) \right\} G(t, a) dS_i(t, a) = 0, \end{aligned} \tag{4.15}$$

where $\bar{A} = \mathcal{S}^{(1)} / \mathcal{S}^{(0)}$. We remind the reader that $\mathcal{S}^{(l)}$ is defined in (4.14).

The proposed score depends on the propensity score $\pi(Z)$, the conditional survival function $S(t|A, Z)$ and the censoring survival function $G(t|A)$. To stress this dependency we use the notation $U_{1,n}^{AIPW}(\beta; \pi, S, G)$ to indicate the score in (4.15).

4.4 Estimation and Inference

4.4.1 Estimation

The proposed score depends on the quantities π, S, G. These quantities, unknown in observational studies, need to be estimated from the data.

We fit the nonparametric Kaplan-Meier estimator to both the treated and the untreated group to obtain a consistent estimator $\hat{G}(t|A)$ such that for $a = 0,1$ it satisfies $\sup_{t\in[0,\tau]} \left|\hat{G}(t \mid a) - G^o(t \mid a)\right|$ $= O_p(n^{-1/2})$. We call $\rho(t,a)$ the influence function such that: $\hat{G}(t \mid a) - G^o(t \mid a) = \frac{1}{n}\sum_{i=1}^n \rho_i(t,a) +$ $o_p(n^{-1/2})$ for $a = 0,1$.

An estimator for the propensity score $\hat{\pi}(Z)$, can be obtained using different methods, parametric or nonparametric, to be chosen by the user. We will use in simulations logistic regression, random forest, support-vector machine and gradient boosted logistic regression.

Estimating $S(t|A,Z)$ is more complicated. The reason being that the Cox model is not collapsible, that is, if the marginal distribution of $T|A$ follows a Cox model, it is not true that the conditional distribution $T|A,Z$ still follows a Cox model and viceversa. This makes the use of classical semiparametric model for estimation of the conditional distribution $T|A,Z$ unsuitable since it would raise compatibility issue with the marginal structural Cox model (4.1) assumed. To overcome this difficulty we propose the use of nonparametric methods such as spline (Gray, 1992; Kooperberg et al., 1995a) and random survival forest (Ishwaran et al., 2008).

Once estimators $\hat{\pi}, \hat{S}, \hat{G}$ are available, we propose to estimate β by solving

$$U_{1,n}^{AIPW}(\beta; \hat{\pi}, \hat{G}, \hat{S}) = 0, \tag{4.16}$$

and $\Lambda_0(t)$ with $\hat{\Lambda}_0(t) = \tilde{\Lambda}_0(t; \hat{\beta}, \hat{\pi}, \hat{S}, \hat{G})$. As usual, we name our estimator $\hat{\beta}$.

4.4.2 Asymptotic properties

We study here the asymptotic properties of our estimators.

Assumption 7. *There exist some functions* $\pi^*(z)$ *and* $S^*(t \mid a, z)$ *such that* $\sup_{z \in \mathcal{Z}} |\hat{\pi}(z) - \pi^*(z)| = O_p(b_n)$*,* $\sup_{t \in [0,\tau], z \in \mathcal{Z}} \left|\hat{S}(t \mid a, z) - S^*(t \mid a, z)\right| = O_p(c_n)$ *for some* $b_n = o(1), c_n = o(1)$ *and* $a = 0, 1$*.*

Assumption 7 is standard in the doubly robust literature, it assumes that the generic estimators $\hat{S}$ and $\hat{\pi}$ converge to working models S^* and π^* that are not necessarily equal to the true quantities (Zhang and Schaubel, 2012b; Yang et al., 2020).

Assumption 8. *There exist two constants* $0 < C_0 < C_1 < 1$*, such that* $0 < C_0 < \inf_{z \in \mathcal{Z}} \pi^*(z) < \sup_{z \in \mathcal{Z}} \pi^*(z) < C_1 < 1$*.*

Assumption 8 is the usual positivity assumption required for IPW based methods (Zhang and Schaubel, 2011, 2012b; Hou et al., 2021).

Assumption 9. *For any* $0 < \pi^*(z) < 1$ *and any survival function* S^**,* $\mathcal{S}^{(0)}(t; \beta^o, \pi^*, S^*, G^o)$ *is strictly positive and there exist bounded* $0 < s^{(l)}(t; \beta^o, \pi^*, S^*, G^o) < \infty$ *such that, for* $l = 0, 1$*:* $\sup_{t \in [0,\tau]} \left|\mathcal{S}^{(l)}(t; \beta^o, \pi^*, S^*, G^o) - s^{(l)}(t; \beta^o, \pi^*, S^*, G^o)\right| = o_p(1)$*.*

Assumption 9 is typical of the Cox model, it can indeed be considered an AIPW version of Assumption B of Andersen and Gill (1982).

The next result proves that, under the assumptions stated above, our proposed estimator $\hat{\beta}$ is consistent if either the propensity score model or the conditional outcome model is correctly specified.

Theorem 10. *Let model* (4.1) *and Assumptions 2-9 hold. Assume* $b_n c_n = o(n^{-1/2})$*, if either* $S^* = S^o$ *or* $\pi^* = \pi^o$*, we have* $\hat{\beta} - \beta^o = o_p(1)$*.*

We now focus on proving asymptotic normality of $\hat{\beta}$. When both models are correctly specified, our estimator is doubly robust in the rate sense. Specifically asymptotic normality of $\hat{\beta}$ does not require $\hat{S}$ or $\hat{\pi}$ to converge to the true S^o or π^o at a $\sqrt{n}$ rate as long as the product of the two rates is $o(\sqrt{n})$. Potentially, one of the two could converge to the true very slowly as long as the other one is fast enough. This property is particularly attractive for our case since we propose the use of machine learning methods like survival random forest for estimation of $S(t|A,Z)$ that are known to have rates of convergence slower than the classical $\sqrt{n}$.

On the other hand, our estimator is also model-doubly robust, that is, it is asymptotically normal as long as one of the two models converge to the true at the classical rate $\sqrt{n}$.

For the next result we need an extra assumption.

Assumption 10. $\mathcal{R} = \{R(t,S^o,G^o) = Y(t) - S^o(t|A,Z)G^o(t|A), \; : \; t \in [0,\tau]\}$ *is a Glivenko-Cantelli class.*

We remind the reader that we indicate with S^o and G^o the true quantities. Assumption 10 is standard in the empirical process literature (Wellner et al., 2013). This assumption is reasonable since $R(t)$ is the difference between an indicator function and a monotone uniformly bounded function.

Theorem 11. *Let model* (4.1) *and Assumptions 2-10 hold. Assume a) or b) or c) below:*

a) (Rate-double robustness): $S^* = S^o$ *and* $\pi^* = \pi^o$ *and* $b_n c_n = o(n^{-1/2})$*,*

or

b) (Model-double robustness): $\pi^* = \pi^o$, $S^* \neq S^o$ *and* $b_n = n^{-1/2}$*; specifically, there exists an influence function* $\phi(z)$ *such that* $\hat{\pi}(z) - \pi^*(z) = \frac{1}{n}\sum_{i=1}^n \phi_i(z) + o_p(n^{-1/2})$*,*

or

c) (Model-double robustness): $S^* = S^o$, $\pi^* \neq \pi^o$ and $c_n = n^{-1/2}$; specifically, there exists an influence function $\psi(t,a,z)$ such that $\hat{S}(t \mid a,z) - S^*(t \mid a,z) = \frac{1}{n}\sum_{i=1}^{n} \psi_i(t,a,z) + o_p(n^{-1/2})$.

We have:

$$\sqrt{n}(\hat{\beta} - \beta^o) = \sigma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_i + o_p(1), \tag{4.17}$$

where explicit forms of σ and $\varphi_i = \varphi_i(\beta^o, \pi^*, S^*, G^o, \phi, \psi, \rho)$ are given in Section 4.8.1 of the Supplementary Material. Therefore, $\sqrt{n}(\hat{\beta} - \beta^o) \rightarrow \mathcal{N}(0, \sigma^{-2} Var(\varphi))$.

When the propensity score is correctly specified (case a) and b) of the Theorem), theory proves that also the Cox-IPW estimator is consistent and asymptotically normal. However, when both models are correct, AIPW is more efficient than IPW (Robins et al., 1995; Van Der Laan et al., 2003). When the conditional outcome model is not correctly specified, no theoretical results exist that compare the efficiency of the AIPW and the IPW estimator. Still, in simulations (Table 4.2, 4.3, 4.5), our estimator shows comparable or better efficiency than Cox-IPW.

The consistency and the asymptotic normality of our estimator does not require specific estimators for the propensity score and the conditional outcome model. The user can therefore choose from a wide variety of estimation techniques as long as the assumptions are satisfied. To the best of our knowledge, little is known on the rate of uniform convergence of nonparametric methods for survival estimation as Survival Random Forest and Spline. Cui et al. (2017) proves that, for fixed covariates, a rate of $n^{-1/(2+d)}$ is achievable by survival random forest, where d is the number of covariates. Kooperberg et al. (1995b) derives instead the L^2 rate of convergence for spline. They show that, under some conditions, the rate can reach $n^{-p/(2p+d)}$, where p is a smoothness parameter; i.e. the optimal global rate for nonparametric regression (Stone, 1982). However, no uniform rates have been provided in the literature. Nevertheless, the rate double robustness of our estimator gives the users the possibility to choose nonparametric estimators for both the estimation of the propensity score and the conditional outcome model. This is a relaxation with respect to Wang and Chen (2001); Tchetgen Tchetgen et al. (2010); Zhang and Schaubel (2012b); Bai et al. (2017); Dukes et al. (2019b); Tan (2019) where classical semiparametric estimators are considered and only the model double robustness of their estimators is proven.

Our estimator is model-doubly robust in the sense that it is both consistent and asymptotically normal even if only one of the two models is correctly specified. In the literature, the concept of model double robustness has been sometimes used only to indicate consistency (Wang and Chen, 2001; Hou et al., 2021).

The complicated expression for φ simplifies if both models are correct and therefore we are under case a) of the Theorem. In this case

$$\begin{aligned}\varphi_i = \varphi_i^a(\beta^o, \pi^o, S^o, G^o) = \int_0^\tau & w_i^o \left\{A_i - \bar{a}(t; \beta^o, \pi^o, S^o, G^o)\right\} \left\{dM_i(t) - E[dM_i(t)|A_i, Z_i]\right\} \\ & + \sum_{a=0,1} \left\{a - \bar{a}(t; \beta^o, \pi^o, S^o, G^o)\right\} E[dM_i(t)|a, Z_i],\end{aligned} \tag{4.18}$$

where $\bar{a} = s^{(1)} \left\{ s^{(0)} \right\}^{-1}$.

Under this specific case we propose a consistent estimator for the asymptotic variance of $\hat{\beta}$.

Theorem 12. *Let model* (4.1) *and Assumptions 2-10 hold. If $S^* = S^o$ and $\pi^* = \pi^o$ with $b_n c_n = o(n^{-1/2})$, the asymptotic variance of $\hat{\beta}$ can be consistently estimated by $\frac{1}{\sqrt{n}} \hat{\sigma}^{-1} \sqrt{\hat{V}}$ where $\hat{V} = \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\varphi}_i^a(\hat{\beta}, \hat{\pi}, \hat{S}, \hat{G}) \right\}^2$ and*

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left\{ \bar{A}^2(t) - \bar{A}(t) \right\} \left[\hat{w}_i dN_i(t) + \hat{w}_i \hat{G}(t, A_i) d\hat{S}_i(t, A_i) - \sum_{a=0,1} \hat{G}(t, a) d\hat{S}_i(t, a) \right],$$

where for simplicity we use $\bar{A}(t)$ to indicate $\bar{A}(t;\hat{\beta},\hat{\pi},\hat{S},\hat{G})$. An explicit expression for $\hat{\varphi}^a$ is given in Section 4.8.1 of the Supplementary Material.

The above result can be used for construction of confidence intervals when both models are correct. When one of the two models is not correct, because of the complexity of the asymptotic variance and because in practice one does not know which model is correct, we suggest the use of nonparametric bootstrap for estimation of the asymptotic variance of $\hat{\beta}$. The use of bootstrap is typical of the literature on double robustness (Zhang and Schaubel, 2012b; Bai et al., 2017; Yang et al., 2020) and the validity of such procedure is due to the fact that $\hat{\beta}$ is asymptotically linear. We will explore both options in simulation. For the second, we will use a standard nonparametric bootstrap, where one draws bootstrap samples from (X_i,δ_i,A_i,Z_i), $i=1,\dots,n$ with replacement.

4.5 Simulations

In this section we study the properties of our estimator on simulated dataset. We compare our AIPW estimator, $\hat{\beta}^{AIPW}$, with the IPW estimator, $\hat{\beta}^{IPW}$, and the naive Cox model that does not adjust for confounders. Moreover, as an oracle estimator we fit the Cox model to the full data.

Simulating data under a marginal structural model is not trivial. Since the covariates Z are not included in the model, it is not straightforward to generate confounding. Following Havercroft and Didelez (2012), we simulate an unobserved variable V that is both associated with the covariates Z and the outcome T inducing confounding. Specific steps of the simulation technique are reported in Section 4.8.6 of the Supplementary Material.

In our simulation we fix $\beta=-1$ and $\lambda_0(t)=1$ and we consider 4 different scenarios explained in Table 4.1. We estimate the propensity score with 4 different methods: logistic regression without interaction, random forest, SVM and boosted logistic regression. For the last three we make use of the following R packages: ranger, e1071 and twang. Unless otherwise specified we use the default settings. Except for random forest, we train our propensity score models in-sample. Since random forest is known to have better out-of-sample performance, we divide the dataset in two dataset with equal size: dat1, dat2. We then fit RF on dat1 to predict the propensity score on dat2 and viceversa. We make use of 3000 trees. We use in simulations stabilized weights, i.e. $w = A \cdot pr(A=1)\{\pi(Z)\}^{-1} + (1-A) \cdot pr(A=0)\{1-\pi(Z)\}^{-1}$ where the marginal $P(A=1)$ is estimated by the empirical proportion.

For the observed outcome model we fit the semiparametric Cox model, survival random forest (Ishwaran et al., 2008) and linear regression spline (Kooperberg et al., 1995a). To this aim we make use of the following R packages: survival, randomForestSRC and polspline.

We simulate 500 dataset with a sample size of 1000. For all scenarios, $41\% - 55\%$ of subjects are treated and $27\% - 33\%$ of subjects are censored. The root of our score is estimated using Newton Raphson routine with 0 as starting value. For IPW the reported standard error is the sandwich estimate of the standard deviation. For AIPW we report as standard error both the model-based estimate defined in Theorem 12, that assumes both models correct, and a bootstrap estimate. For the latter, 50 bootstrap samples are used to save computational time.

Results of simulations 1-4 are reported in Table 4.2-4.5, respectively. In Scenarios 1 and 2, AIPW outperforms IPW in term of estimation of both the treatment effect β and the asymptotic variance of the estimator. In Scenario 3, IPW estimates are biased. Our method instead shows consistency exhibiting protection against the misspecified propensity scores. As expected, the AIPW model-based standard error underestimates the empirical standard deviation, leading to confidence intervals with coverage below the nominal 95%. However, bootstrap confidence intervals show nominal coverage. In Scenario 4, when the treatment model is assumed to be of logistic form, $\hat{\beta}^{IPW}$ shows some bias, bias that is corrected by $\hat{\beta}^{AIPW}$. For the other propensity scores, $\hat{\beta}^{AIPW}$ bias is always smaller or comparable to $\hat{\beta}^{IPW}$ one. However, as in Scenarios 1 and 2, the model-based standard error of $\hat{\beta}^{AIPW}$ outperforms the IPW sandwich estimator.

By theory, the model-based variance estimator is inconsistent when only one of the two models is correctly specified. This is reflected in Scenario 3, where the propensity score is misspecified. In the other scenarios, however, the model-based variance estimator shows some level of robustness. On the other hand, as expected, the bootstrap confidence intervals show good coverage in all 4 scenarios. This is in line with Funk et al. (2011).

To our best knowledge little experience exists in the literature to inform us how to properly tune in practice survival random forest. Moreover no valuable softwares exist to find the optimal hyperparameters. In Scenarios 3 and 4 we tune SRF paying particular attention to the split rule, the depth of the node (nodedepth), the size of the terminal nodes (nodesize), the number of trees (ntree) and the number of variables randomly selected as candidates for splitting a node (mtry). In Scenario 3 we fit SRF with nodesize=15, ntree=2000, mtry=4, nsplit=5, split rule=bsgradient. In Scenario 4 we fit SRF with nodesize=8, ntree=2000, mtry=7, nsplit=2, nodedepth=10, split rule=bsgradient. However, we do not try all the possible combinations choosing a set of hyperparameters that might not be the optimal. In both scenarios, spline outperforms SRF in term of performance. Our simulations seem to suggest that while the latter needs to be properly tuned, Spline has good performance when the default hyperparameters are used.

Table 4.1: Data-generating mechanisms of Scenarios 1-4. $\beta = -1$ and $\Lambda_0(t) = t$ are fixed. In Scenarios 1-3, $Z = [Z_1, Z_2, Z_3]^\top$. In Scenario 4, $Z = [Z_1, Z_2, Z_3, Z_4, Z_5, Z_6]^\top$.

Scenario	Data-generating mechanism
1 PS: Logistic	$V \sim U(0,1)$ $Z_1 = 0.5V + \mathcal{N}(0,0.5), Z_2 = 0.3V + \mathcal{N}(0,1), Z_3 = V^2 + \mathcal{N}(0,0.3)$ $\text{logit}\{\pi(Z)\} = -Z_1 + Z_2 - Z_3$ $C(a) \sim Exp(1/5 + 1/5a)$
2 PS: Logistic with Interaction	$V \sim U(0,1)$ $Z_1 = 0.5V + U(-0.5,0.5), Z_2 = 0.3V + B(0.5), Z_3 = V^2 + U(-0.3,0.3)$ $\text{logit}\{\pi(Z)\} = -Z_1 + Z_2 + Z_3 + Z_1Z_2 - Z_2Z_3 + Z_1Z_2Z_3$ $C(a) \sim Exp(1/8 + 1/8a)$
3 PS: Soft Partition 1	$V \sim U(0,1)$ $Z_1 = 0.5V + \mathcal{N}(0,0.5), Z_2 = 0.3V + \mathcal{N}(0,1), Z_3 = 0.1V + \mathcal{N}(0,0.3)$ $\varepsilon = B(\text{expit}(-Z_1 + Z_2 + Z_3 + Z_1Z_2 - Z_2Z_3 + Z_1Z_2Z_3))$ $\pi(Z) = \mathbf{1}\{Z_1 + Z_2 + Z_3 + \varepsilon < 0.5\}$ $C(a) \sim Exp(1/10 + 1/10a)$
4 PS: Soft Partition 2	$V \sim N(0,1)$ $Z_1, Z_2, Z_3, Z_4, Z_5, Z_6 = (V + N(0,1))/\sqrt{2}$ $\pi(Z) = 0.8 * \mathbf{1}\{\sum_{i=1}^{6} Z_i^2 < \chi_{0.5,6}\} + 0.2 * \mathbf{1}\{\sum_{i=1}^{6} Z_i^2 > \chi_{0.5,6}\}$ $C(a) \sim Exp(1/8 + 1/8a)$

Table 4.2: Results of simulations from Scenario 1. The true $\beta = -1$. Column PS and OC indicates how the propensity score and the conditional outcome model are estimated, respectively. The first SE and CP are model-based. The second SE and CP are based on bootstrap. Bootstrap is performed only for the first 100 simulations. SD, standard deviation; SE, standard error; CP, coverage of a 95% confidence interval; Boot, bootstrap.

	IPW					**AIPW**			
PS	Bias	SD	SE	CP	**OC**	Bias	SD	SE	CP
								Model / Boot	Model / Boot
Log	−0.016	0.090	0.205	1	Cox	−0.017	0.127	0.111 / 0.109	0.96 / 0.98
					SRF	−0.002	0.099	0.082 / 0.090	0.95 / 0.94
					Spline	−0.011	0.090	0.089 / 0.097	0.97 / 0.95
RF	−0.016	0.111	0.305	1	Cox	−0.020	0.117	0.133 / 0.118	0.98 / 0.97
					SRF	−0.014	0.109	0.179 / 0.105	0.96 / 0.95
					Spline	−0.011	0.111	0.115 / 0.113	0.97 / 0.97
SVM	−0.100	0.088	0.122	0.95	Cox	0.004	0.078	0.079 / 0.084	0.95 / 0.96
					SRF	−0.004	0.077	0.074 / 0.078	0.94 / 0.97
					Spline	−0.002	0.077	0.078 / 0.085	0.95 / 0.96
Tw	−0.099	0.083	0.119	0.93	Cox	0.018	0.077	0.072 / 0.073	0.92 / 0.94
					SRF	0.0005	0.077	0.068 / 0.074	0.91 /0.94
					Spline	0.001	0.077	0.072 / 0.081	0.93 / 0.94
	Oracle					**Naive Cox**			
	−0.003	0.028	0.028	0.94		−0.302	0.086	0.084	0.13

Table 4.3: Results of simulations from Scenario 2. The true $\beta = -1$. Column PS and OC indicates how the propensity score and the conditional outcome model are estimated, respectively. The first SE and CP are model-based. The second SE and CP are based on bootstrap. Bootstrap is performed only for the first 100 simulations. SD, standard deviation; SE, standard error; CP, coverage of a 95% confidence interval; Boot, bootstrap.

	IPW					**AIPW**			
PS	Bias	SD	SE	CP	**OC**	Bias	SD	SE	CP
								Model / Boot	Model / Boot
					Cox	0.004	0.065	0.068 / 0.070	0.96 / 0.96
Log	−0.010	0.061	0.108	1	SRF	−0.002	0.054	0.052 / 0.057	0.94 / 0.95
					Spline	0.003	0.058	0.059 / 0.078	0.95 / 0.97
					Cox	−0.041	0.097	0.121 / 0.103	0.99 / 0.95
RF	0.056	0.091	0.187	1	SRF	−0.002	0.084	0.077 / 0.082	0.95 / 0.94
					Spline	−0.015	0.097	0.098 / 0.079	0.94 / 0.91
					Cox	0.049	0.062	0.061 / 0.063	0.86 / 0.85
SVM	−0.142	0.077	0.098	0.73	SRF	0.0003	0.053	0.050 / 0.052	0.93 / 0.96
					Spline	0.012	0.063	0.056 / 0.074	0.91 / 0.96
					Cox	0.031	0.052	0.058 / 0.050	0.94 / 0.88
Tw	−0.083	0.055	0.093	0.96	SRF	−0.011	0.049	0.047 / 0.047	0.94 / 0.93
					Spline	0.005	0.052	0.052 / 0.059	0.94 / 0.97
	Oracle					**Naive Cox**			
	−0.001	0.027	0.028	0.95		−0.322	0.080	0.085	0.02

Table 4.4: Results of simulations from Scenario 3. The true $\beta = -1$. Column PS and OC indicates how the propensity score and the conditional outcome model are estimated, respectively. The first SE and CP are model-based. The second SE and CP are based on bootstrap. Bootstrap is performed only for the first 100 simulations. SD, standard deviation; SE, standard error; CP, coverage of a 95% confidence interval; Boot, bootstrap.

	IPW					**AIPW**			
PS	Bias	SD	SE	CP	**OC**	Bias	SD	SE	CP
								Model / Boot	Model / Boot
Log	−0.993	0.133	0.261	0.04	Cox	0.065	0.150	0.023 / 0.102	0.58 / 0.86
					SRF	−0.040	0.140	0.052 / 0.121	0.55 / 0.97
					Spline	−0.003	0.135	0.057 / 0.150	0.56 / 0.95
RF	−1.030	0.149	0.314	0.12	Cox	0.063	0.108	0.057 / 0.100	0.58 / 0.90
					SRF	−0.053	0.144	0.056 / 0.125	0.57 / 0.94
					Spline	−0.006	0.138	0.061 / 0.328	0.60 / 0.98
SVM	−0.970	0.143	0.374	0.19	Cox	0.064	0.108	0.063 / 0.106	0.64 / 0.86
					SRF	−0.028	0.142	0.062 / 0.124	0.60 / 0.98
					Spline	−0.001	0.137	0.068 / 0.256	0.65 / 0.97
Tw	−1.142	0.125	0.174	0	Cox	0.064	0.104	0.042 / 0.097	0.47 / 0.89
					SRF	−0.064	0.142	0.041 / 0.120	0.44 / 0.96
					Spline	−0.009	0.137	0.045 / 0.284	0.47 / 0.96
	Oracle					**Naive Cox**			
	−0.001	0.029	0.028	0.93		−1.577	0.112	0.129	0

Table 4.5: Results of simulations for Scenario 4. The true $\beta = -1$. Column PS and OC indicates how the propensity score and the conditional outcome model are estimated, respectively. The first SE and CP are model-based. The second SE and CP are based on bootstrap. Bootstrap is performed only for the first 100 simulations. SD, standard deviation; SE, standard error; CP, coverage of a 95% confidence interval; Boot, bootstrap.

	IPW					**AIPW**			
PS	Bias	SD	SE	CP	**OC**	Bias	SD	SE	CP
								Model / Boot	Model / Boot
Log	0.159	0.054	0.080	0.50	Cox	−0.106	0.139	0.062 / 0.170	0.81 / 0.91
					SRF	0.029	0.056	0.044 / 0.097	0.89 / 0.91
					Spline	0.004	0.049	0.054 / 0.057	0.97 / 0.97
RF	−0.006	0.070	0.212	1	Cox	0.005	0.066	0.109 / 0.058	0.97 / 0.91
					SRF	0.003	0.061	0.048 / 0.065	0.89 / 0.95
					Spline	0.004	0.059	0.048 / 0.059	0.95 / 0.93
SVM	0.003	0.073	0.105	1	Cox	−0.001	0.049	0.048 / 0.049	0.95 / 0.96
					SRF	0.020	0.053	0.048 / 0.053	0.88 / 0.92
					Spline	0.004	0.049	0.048 / 0.049	0.96 / 0.97
Tw	0.025	0.057	0.088	1	Cox	−0.008	0.047	0.045 / 0.045	0.94 / 0.93
					SRF	0.032	0.054	0.043 / 0.053	0.81 / 0.86
					Spline	0.005	0.047	0.044 / 0.046	0.94 / 0.95
	Oracle					**Naive Cox**			
	−0.002	0.028	0.028	0.95		0.162	0.077	0.084	0.54

4.6 Real Data

We study the effect of mid-life alcohol consumption on overall death. To this aim we use data from the Honolulu-Asia Aging Study (HAAS). The study, established in 1991 as a continuation of the Honolulu Hearth Program project (HHP), collected data on a cohort of Japanese men with a focus on causes of cognitive and motor impairment, stroke, and the common chronic conditions of late-life.

The mid-life alcohol exposure was assessed at exam 1 and exam 3 of HHP (1965/ 1971-1974) by self report. People with a light exposure to alcohol at both exams are considered light drinkers, while people with a heavy exposure to alcohol in at least one of the two life periods are considered heavy drinkers. The death of the participants, when available, was collected from their death certificates.

To study the effect of alcohol exposure on overall survival it is important to adjust for confounding. To this aim we use as covariates age at baseline, maximum years of education, ApoE genotype, systolic blood pressure, and heart rate. The summary statistics of these variables can be found in Table 4.7 in Section 4.8.7 of the Supplementary Material.

Since HAAS starts at exam 4 (1991), we consider exam 4 as time 0 and we restrict the analysis to the set of participants still available. After eliminating some observations (~ 50) with missing entries we are left with 2061 participants; 1509 light drinkers and 552 heavy drinkers.

Among light drinkers 1317 (87%) deaths were observed while among heavy drinkers 506 (92%) deaths were recorded. The Kaplan-Meier curves for the two groups are presented in Figure 4.2 in Section 4.8.7 of the Supplementary Material.

We use our proposed score to estimate the effect of mid-life alcohol exposure on overall survival. As in simulations we estimate the propensity score by logistic regression without interaction, random forest, SVM with sigmoid kernel and boosted logistic regression (twang). In Figure 4.1 we plot the distribution of the estimated propensity scores for both groups. We estimate the conditional outcome model by Cox model, survival random forest and spline. For SRF we use 500 trees, we set the terminal nodes' size at 30, the number of variables randomly selected as candidates for splitting a node at 5 and the splitting rule at bs.gradient. For spline we use a penalty of 0.5.

The results of the analysis are reported in Table 4.6. For comparison we also report the results of the naive Cox model that doesn't adjust for confounding and IPW.

In line with Figure 4.2 all the results seem to suggest that mid-life alcohol exposure has a significant effect on overall survival with a positive hazard ratio between heavy and light drinkers.

The magnitude of the effect changes according to the method used. AIPW with Cox estimate ranges between 0.243 and 0.256. AIPW with SRF estimate ranges between 0.224 and 0.255. While when spline is used the estimated effect ranges between 0.240 and 0.259. For IPW the estimated effect ranges from 0.242 and 0.283, while the naive Cox model gives a point estimate of 0.282. As expected, AIPW confidence intervals provide a better representation of the causal effect. Both the naive Cox and the IPW with SVM estimates of the treatment effect are around 0.283 suggesting that SVM model for propensity score might not properly adjust for confounding. However, AIPW estimates are stable across the different propensity scores corroborating their robustness with respect to the estimation of the propensity score.

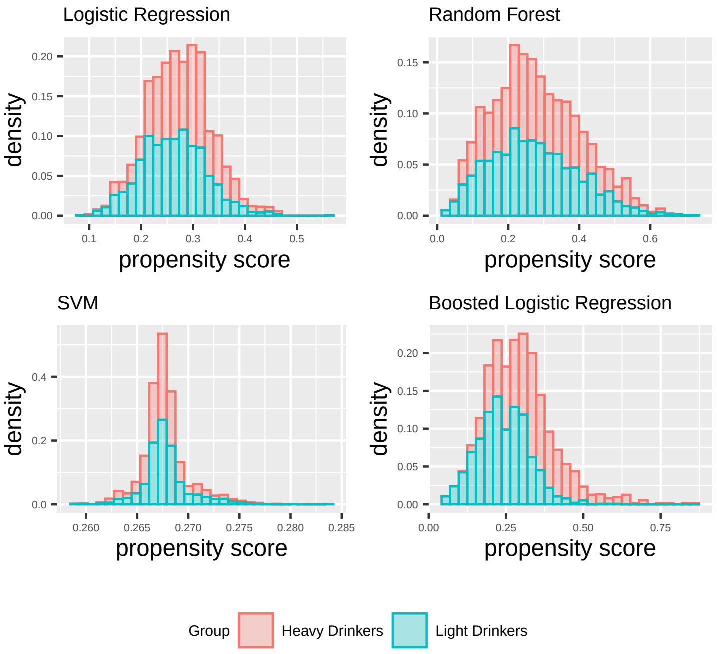


Figure 4.1: Distribution of the estimated propensity score for the HHP-HAAS dataset.

Table 4.6: Estimated treatment effect for the HHP-HAAS dataset. Column PS and OC indicates how the propensity score and the conditional outcome model are estimated, respectively. The first CI and p-value are model-based while the second are based on bootstrap. The computed P-value is two-sided. CI, confidence interval; Boot; bootstrap.

	IPW				**AIPW**		
PS	$\hat{\beta}$	CI	P-value	OC	$\hat{\beta}$	CI	P-value
						Model / Boot	Model / Boot
Log	0.251	$[0.133, 0.368]$	< 0.001	Cox	0.243	$[0.142, 0.345]$ / $[0.144, 0.343]$	< 0.001 / < 0.001
				SRF	0.240	$[0.143, 0.338]$ / $[0.146, 0.335]$	< 0.001 / < 0.001
				Spline	0.243	$[0.141, 0.345]$ / $[0.139, 0.348]$	< 0.001 / < 0.001
RF	0.242	$[0.078, 0.406]$	0.004	Cox	0.245	$[0.123, 0.368]$ / $[0.080, 0.410]$	< 0.001 / 0.004
				SRF	0.255	$[0.134, 0.376]$ / $[0.109, 0.402]$	< 0.001 / < 0.001
				Spline	0.242	$[0.120, 0.364]$ / $[0.075, 0.409]$	< 0.001 / 0.005
SVM	0.283	$[0.182, 0.384]$	< 0.001	Cox	0.256	$[0.161, 0.352]$ / $[0.160, 0.353]$	< 0.001 / < 0.001
				SRF	0.250	$[0.158, 0.342]$ / $[0.157, 0.342]$	< 0.001 / < 0.001
				Spline	0.259	$[0.164, 0.354]$ / $[0.149, 0.369]$	< 0.001 / < 0.001
Tw	0.245	$[0.144, 0.345]$	< 0.001	Cox	0.243	$[0.152, 0.333]$ / $[0.152, 0.333]$	< 0.001 / < 0.001
				SRF	0.224	$[0.137, 0.311]$ / $[0.138, 0.310]$	< 0.001 / < 0.001
				Spline	0.240	$[0.150, 0.331]$ / $[0.142, 0.339]$	< 0.001 / < 0.001
	Naive Cox						
	0.282	$[0.177, 0.388]$	< 0.00001				

4.7 Discussion

We have derived a new score for the estimation of the causal hazard ratio. Our proposal is doubly robust with respect to the propensity score and the conditional survival function of the failure time. Our score augments the Cox-IPW score to protect against possible misspecification of the propensity score.

The potential censoring times are assumed to be independent of both the potential failure times and the confounders. This assumption could be relaxed following AIPWCC methodology (Rotnitzky and Robins, 2005; Tsiatis, 2007; Bai et al., 2017; Zhang and Schaubel, 2012b). Under the weaker assumption of independence between the potential failure times and the potential censoring times given the confounders, AIPWCC methodology treats both censoring and treatment indicator as coarsening variables. The IPWCC score weights every observation by the inverse of the probability of receiving the observed treatment and of being uncensored. The AIPWCC score augments the IPWCC to protect against possible misspecification of both the propensity score and the censoring mechanism. The AIPWCC score would then have an extra term with respect to our proposed score. Moreover it would be doubly robust with respect to the models corresponding to the weights, propensity score and censoring distribution, and the conditional survival function of the failure time. The computation of the projection that defines the augmentation for the IPWCC score is non trivial; such score is beyond the scope of this work and we leave it for future works.

We have proved that our score is both model and rate-doubly robust. As explained, the latter characteristic allows the user to choose from a variety of methodologies, both parametric and nonparametric, for estimation of the propensity score and the outcome model. In simulations we have investigated the performance of different nonparametric methods such as random forest, SVM, boosted logistic regression for the estimation of the propensity score and survival random forest and spline for the estimation of the conditional survival function of the failure time. The idea that nonparametric methods are always consistent, because in principle model-free is a common misconception. While it is true that they relax the modeling assumptions typical of parametric methodologies, their consistency is not granted and it is often hard to assess. Moreover, they have often convergence rates slower than the classical $\sqrt{n}$. It is therefore convenient to pair them with estimators that are both model and rate-doubly robust as our proposal. The tuning process of nonparametric methods can be non trivial. In simulations we have discovered how in practice tuning survival random forest can be quite complicated and how the default settings are not always optimal. On the other hand Spline has shown good performance with the default parametrization.

Because of the non-collapsibility of the hazard ratio, assuming that the conditional distribution of the failure time follows a Cox model is incompatible with the marginal structural Cox model. However, in simulations we have investigated the performance of our estimator when the conditional outcome model is assumed to follow the Cox model. Even though using survival random forest and spline always outperformed the use of the Cox model, the latter has still been proven useful by our simulations to correct for mistakes in the estimation of the propensity score. This perhaps comes with no surprise since experience has shown that Cox model is quite robust to model misspecification for survival prediction. We however advise the user to use nonparametric methods for estimation of the outcome model.

In simulations we have compared our score with the state-of-the-art Cox IPW. To estimate its asymptotic variance we have used the proposed sandwich estimator. This estimator does not take into account the weights' estimation and it is therefore known to be slightly biased. Practitioners have been recently used bootstrap instead. However we have shown in simulations that our model-based variance estimator has better performance than the IPW sandwich estimator.

The ideas beyond the derivation of the AIPW score described in section 4.3 are not necessarily exclusive to the marginal structural Cox model; this work opens up a new line of research.

4.8 Appendix

4.8.1 Formal quantities

In Theorem 11 we claim that $\sqrt{n}(\hat{\beta}-\beta^o) = \sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i + o_p(1)$. We report here the expressions for φ and σ.

$$
\begin{aligned}
&\varphi_i(\beta^o,\pi^*,S^*,G^o,\phi,\psi,\rho)\\
&=\int_0^\tau w_i^*\left\{A_i-\bar{a}(t;\beta^o,\pi^*,S^*,G^o)\right\}\\
&\quad\times\left\{dM_i(t)+dS_i^*(t,A_i)G^o(t,A_i)+e^{\beta^o A_i}S_i^*(t,A_i)G^o(t,A_i)d\Lambda_0^o(t)\right\}\\
&\quad-\int_0^\tau\sum_{a=0,1}\left\{a-\bar{a}(t;\beta^o,\pi^*,S^*,G^o)\right\}\left\{dS_i^*(t,a)G^o(t,a)+e^{\beta^o a}S_i^*(t,a)G^o(t,a)d\Lambda_0^o(t)\right\}\\
&\quad-\int_0^\tau\frac{1}{n}\sum_{j=1}^{n}\phi_i(Z_j)\left\{\pi_j^*\right\}^{-2}\left(dN_j(t)+G^o(t,A_j)dS_j^*(t,A_i)-d\Lambda_0^o(t)e^{\beta^o A_j}R_j(t,S^*,G^o)\right) &&(4.19)\\
&\quad-\int_0^\tau\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}B_i(t,\beta^o,\pi^*,S^*,G^o) &&(4.20)\\
&\quad-\int_0^\tau\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}D_i(t,\beta^o,\pi^*,S^*,G^o) &&(4.21)\\
&\quad+\int_0^\tau d\Lambda_0^o(t)\frac{1}{n}\sum_{j=1}^{n}\left\{G^o(t,1)\psi_i(t,1,Z_j)+S_j^*(t,1)\rho_i(t,1)\right\}\left\{\frac{A_j}{\pi_j^*}e^{\beta^o A_j}-e^{\beta^o}\right\}\\
&\quad+\int_0^\tau\frac{1}{n}\sum_{j=1}^{n}\left\{G^o(t,A_j)d\psi_i(t,1,Z_j)+dS_j^*(t,A_j)\rho_i(t,1)\right\}\left\{\frac{A_j}{\pi_j^*}-1\right\}, &&(4.22)
\end{aligned}
$$

where

$$
\begin{aligned}
&B_i(t,\beta^o,\pi^*,S^*,G^o)\\
&=\left\{ s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)-\frac{1}{n^2}\sum_{l,m=1}^{n}\phi_l(Z_m)\{\pi_m^*\}^{-2}A_m e^{\beta^o A_m}R_m(t,S^*,G^o)\right\}\\
&\qquad\times\frac{1}{n}\sum_{j=1}^{n}\phi_i(Z_j)\left[\frac{1-A_j}{\left\{1-\pi_j^*\right\}^2}-\frac{A_j}{\left\{\pi_j^*\right\}^2}\right]\\
&\qquad\times\left\{dN_j(t)+G^o(t,A_j)dS_j^*(t,A_j)-d\Lambda_0^o(t)\exp(\beta^o A_j)R_j(t,S^*,G^o)\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
D_i(t,\beta^o,\pi^*,S^*,G^o)=&\left[s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)+\frac{1}{n}\sum_{l=1}^{n}J_l(t,1)\left\{\frac{A_l}{\pi_l^*}e^{\beta^o A_l}-e^{\beta^o}\right\}\right]\\
&\times\frac{1}{n}\sum_{j=1}^{n}\Big(w_j^*\left\{G^o(t,A_j)d\psi_i(t,A_j,Z_j)+dS_j^*(t,A_j)\rho_i(t,A_j)\right\}\\
&\quad-\sum_{a=0,1}\left\{G^o(t,a)d\psi_i(t,a,Z_j)+dS_j^*(t,a)\rho_i(t,a)\right\}\\
&\quad+d\Lambda_0^o(u)\Big\{w_j^*e^{\beta^o A_j}\left\{G^o(t,A_j)\psi_i(t,A_j,Z_j)+S_j^*(t,A_j)\rho_i(t,A_j)\right\}\\
&\quad-\sum_{a=0,1}e^{\beta^o a}\left\{G^o(t,a)\psi_i(t,a,Z_j)+S_j^*(t,a)\rho_i(t,a)\right\}\Big).
\end{aligned}
$$

Moreover

$$
\sigma=\int_0^\tau\left\{\bar{a}^2(t;\beta^o,\pi^*,S^*,G^o)-\bar{a}(t;\beta^o,\pi^*,S^*,G^o)\right\}d\Lambda_0^o(t)s^{(0)}(t;\beta^o,\pi^*,S^*,G^o).
$$

The complicated expression of φ_i simplifies according to which model is correctly specified. Specifically, if both models are correct (case a) of Theorem 11), lines (4.19)-(4.22) are negligeable. On the other hand, if only the propensity score model is correct (case b) of Theorem 11), lines

(4.21)-(4.22) are negligeable and if only the conditional outcome model is correct (case c) of Theorem 11), lines (4.19)-(4.20) are negligeable.

In Theorem 12, when both models are correct we derive the following consistent estimator for φ:

$$\begin{aligned}\hat{\varphi}_i^a(\hat{\beta},\hat{\pi},\hat{S},\hat{G}) = \int_0^\tau & \hat{w}_i \left\{A_i - \bar{A}(t;\hat{\beta},\hat{\pi},\hat{S},\hat{G})\right\} \\ & \times \left\{dM_i(t;\hat{\beta},\hat{\Lambda}_0) + d\hat{S}_i(t,A_i)\hat{G}(t,A_i) + d\hat{\Lambda}_0^o(t)\hat{S}_i(t,A_i)\hat{G}(t,A_i)\right\} \\ & - \int_0^\tau \sum_{a=0,1} \left\{a - \bar{A}(t;\hat{\beta},\hat{\pi},\hat{S},\hat{G})\right\} \left\{d\hat{S}_i(t,a)\hat{G}(t,a) + d\hat{\Lambda}_0(t)\hat{S}_i(t,a)\hat{G}(t,a)\right\},\end{aligned}$$

where

$$M_i(t;\hat{\beta},\hat{\Lambda}_0) = N_i(t) - Y_i(t)e^{\hat{\beta}A_i}\hat{\Lambda}_0(t).$$

4.8.2 Proof of the main results

We remind the reader that we use the following notation:

$$\begin{aligned}U_{1,n}^{AIPW}(\beta;\pi,S,G) = \int_0^\tau \frac{1}{n}\sum_{i=1}^n & w_i \left\{A_i - \bar{A}(t;\beta,\pi,S,G)\right\}\left\{dN_i(t) + G(t,A_i)dS_i(t,A_i)\right\} \\ & - \sum_{a=0,1}\left\{a - \bar{A}(t;\beta,\pi,S,G)\right\} G(t,a)dS_i(t,a),\end{aligned}$$

where $w = [A\pi(Z) + (1-A)\{1-\pi(Z)\}]^{-1}$.

Moreover we remind the reader that the above is equivalent to:

$$\begin{aligned}U_{1,n}^{AIPW}(\beta,\pi,S,G) = \int_0^\tau \frac{1}{n}\sum_{i=1}^n [&w_i A_i \{dN_i(t)+G(t,A_i)dS_i(t,A_i)\\ &-d\tilde{\Lambda}_0(t;\beta,\pi,S,G)\,e^{\beta A_i}R_i(t,S,G)\Big\}\\ &-\Big\{G(t,1)dS_i(t,1)+d\tilde{\Lambda}_0(t;\beta,\pi,S,G)\,e^{\beta}G(t,1)S_i(t,1)\Big\}\Big]\end{aligned}$$

where $\tilde{\Lambda}_0(t;\beta,\pi,S,G)$ is the solution to $U_{2,n}^{AIPW}(t,\Lambda;\beta,\pi,S,G)=0$ where:

$$\begin{aligned}&U_{2,n}^{AIPW}(t,\Lambda_0;\beta,\pi,S,G)\\ &=\frac{1}{n}\sum_{i=1}^n\int_0^\tau\Bigg[w_i\Big\{dN_i(u)+dS_i(u,A_i)G(u,A_i)-d\Lambda_0(u)e^{\beta A_i}R_i(u,S,G)\Big\}\\ &\quad-\sum_{a=0,1}dS_i(u,a)G(u,a)-d\Lambda_0(u)\sum_{a=0,1}e^{\beta a}S_i(u,a)G(u,a)\Bigg].\end{aligned}$$

Proof of Lemma 20

For this proof we make use of some additional lemmas, Lemmas 24-26, reported in Section 4.8.5.

Proof of Lemma 20. We start proving that, for a generic function of the observed data $q(X,\delta,A,Z)$:

$$\prod\{q(X,\delta,A,Z)\mid\mathcal{T}\}=\mathrm{E}\{q(X,\delta,A,Z)\mid A,Z\}-\mathrm{E}\{q(X,\delta,A,Z)\mid Z\}. \tag{4.23}$$

Define $\mathcal{T}^{'}=\{\phi(A,Z)\ \ for\ all\ \phi\}$. By Lemma 24, $\mathcal{T}\subset\mathcal{T}^{'}$ and hence by Lemma 25, we have:

$$\begin{aligned}\prod\{q(X,\delta,A,Z)\mid\mathcal{T}\}&=\prod\Big[\prod\Big\{q(X,\delta,A,Z)|\mathcal{T}^{'}\Big\}\Big|\mathcal{T}\Big]\\ &=\prod[\mathrm{E}\{q(X,\delta,A,Z)|A,Z\}|\mathcal{T}].\end{aligned}$$

Now, by Lemma 26, the above equals to:

$$= \mathrm{E}\{q(X,\delta,A,Z) \mid A,Z\} - \mathrm{E}\left[\mathrm{E}\{q(X,\delta,A,Z) \mid A,Z\} \mid Z\right]$$
$$= \mathrm{E}\{q(X,\delta,A,Z) \mid A,Z\} - \mathrm{E}\{q(X,\delta,A,Z) \mid Z\},$$

where the last line comes from an application of the tower law of conditional expectation.

We now apply (4.23) to our specific U^{IPW}. We have:

$$\begin{aligned} \prod\left\{U_1^{IPW}|\mathcal{T}\right\} &= \mathrm{E}\left\{U_1^{IPW} \mid A,Z\right\} - \mathrm{E}\left\{U_1^{IPW} \mid Z\right\} \\ &= \int_0^\tau \left[\mathrm{E}\{wAdM(t) \mid A,Z\} - \mathrm{E}\{wAdM(t) \mid Z\}\right]. \end{aligned} \tag{4.24}$$

Now, calculating the second conditional expectation, the above equals:

$$= \int_0^\tau \left[\mathrm{E}\{wAdM(t) \mid A,Z\} - \sum_{a=0,1} wP(A=a|Z)\mathrm{E}\{AdM(t) \mid A=a,Z\}\right]$$
$$= \int_0^\tau \left[wAd\mathrm{E}\{M(t) \mid A,Z\} - \mathrm{E}\{dM(t) \mid A=1,Z\}\right].$$

Similarly:

$$\begin{aligned} \prod\left\{U_2^{IPW}(t) \mid \mathcal{T}\right\} &= \mathrm{E}\left\{U_2^{IPW}(t) \mid A,Z\right\} - \mathrm{E}\left\{U_2^{IPW}(t) \mid Z\right\} \\ &= \int_0^t \left[\mathrm{E}\{wdM(u) \mid A,Z\} - \mathrm{E}\{wdM(u) \mid Z\}\right] \\ &= \int_0^t \left[wd\mathrm{E}\{M(u) \mid A,Z\} - \mathrm{E}\{dM(u) \mid A=1,Z\} - \mathrm{E}\{dM(u) \mid A=0,Z\}\right]. \end{aligned} \tag{4.25}$$

The result of the Lemma follows directly from (4.24) and (4.25). □

Proof of Theorem 9

Proof of Theorem 9. We prove separately that $\mathrm{E}\left\{U_1^{AIPW}(\beta^o,\Lambda_0^o;\pi,S,G^o)\right\}=0$ and $\mathrm{E}\left\{U_2^{AIPW}(t,\beta^o,\Lambda_0^o;\pi,S,G^o)\right\}=0$ for each $t\in[0,\tau]$ if either $\pi=\pi^o$ and $S=S^o$.

If $\pi=\pi^o$, we have:

$$
\begin{aligned}
&\mathrm{E}\left\{U_1^{AIPW}(\beta^o,\Lambda_0^o;\pi^o,S,G^o)\right\}\\
&\quad=\int_0^\tau \mathrm{E}\left[\frac{\mathrm{E}\{AdM(t)|Z\}}{\pi^o(Z)}\right]\\
&\quad\quad+\int_0^\tau \mathrm{E}\left(\frac{1}{\pi^o(Z)}\mathrm{E}\left[A\left\{G^o(t|A)dS(t|A,Z)+d\Lambda_0^o(t)e^{\beta^o A}G^o(t|A)S(t|A,Z)\right\}|Z\right]\right)\\
&\quad\quad-\int_0^\tau \mathrm{E}\left[G^o(t|1)dS(t|1,Z)-d\Lambda_0^o(t)e^{\beta^o}G^o(t|1)S(t|1,Z)\right].
\end{aligned}
$$

Calculating the conditional expectation in the above equation we have:

$$
\begin{aligned}
&\mathrm{E}\left\{U_1^{AIPW}(\beta^o,\Lambda_0^o;\pi^o,S,G^o)\right\}\\
&\quad=\int_0^\tau \mathrm{E}\left[dM^1(t)\right]+\mathrm{E}\left\{G^o(t|1)dS(t|1,Z)+d\Lambda_0^o(t)e^{\beta^o}G^o(t|1)S(t|1,Z)\right\}\\
&\quad\quad-\int_0^\tau \mathrm{E}\left\{G^o(t|1)dS(t|1,Z)+d\Lambda_0^o(t)e^{\beta^o}G^o(t|1)S(t|1,Z)\right\}\\
&\quad=\int_0^\tau \mathrm{E}\left[dM^1(t)\right]=0.
\end{aligned}
$$

On the other hand, if $S=S^o$, since by (4.11)

$$
\int_0^\tau \mathrm{E}\{dM(t)|A,Z\}=-\int_0^\tau G^o(t|A)dS^o(t|A,Z)-\int_0^\tau d\Lambda_0^o(t)\exp(\beta^o A)G^o(t|A)S^o(t|A,Z),
$$

we have:

$$
\begin{aligned}
&\mathrm{E}\left\{U_1^{AIPW}(\beta^o,\Lambda_0^o,\pi,S^o,G^o)\right\} \\
&\quad = \int_0^\tau \mathrm{E}\left[\frac{A\mathrm{E}\{dM(t)|A,Z\}}{\pi(Z)}\right] - \int_0^\tau \mathrm{E}\left[\frac{A\mathrm{E}\{dM(t)|A,Z\}}{\pi(Z)}\right] + \int_0^\tau \mathrm{E}\left[\mathrm{E}\{dM(t)|A=1,Z\}\right] \\
&\quad = \int_0^\tau \mathrm{E}\left[\mathrm{E}\{dM(t)|A=1,Z\}\right] = \int_0^\tau \mathrm{E}\left[dM^1(t)\right] = 0.
\end{aligned}
$$

Therefore, if either $\pi = \pi^o$ or $S = S^o$, we have:

$$
\mathrm{E}\left\{U_1^{AIPW}(\beta^o,\Lambda_0^o;\pi,S,G^o)\right\} = \int_0^\tau \mathrm{E}\left\{dM^1(t)\right\} = 0.
$$

Similarly to before, we have for $t \in [0,\tau]$, if $\pi = \pi^o$:

$$
\begin{aligned}
&\mathrm{E}\left\{U_2^{AIPW}(t,\beta^o,\Lambda_0^o;\pi^o,S,G^o)\right\} = \int_0^t \mathrm{E}\left[w^o\mathrm{E}\{dM(u)|Z\}\right] \\
&\quad - \sum_{a=0,1}\int_0^t \mathrm{E}\left\{G^o(u|a)dS(u|a,Z) - d\Lambda_0^o(u)e^{\beta^o a}G^o(u|a)S(u|a,Z)\right\} \\
&\qquad + \int_0^t \mathrm{E}\left(\mathrm{E}\left[w^o\left\{G^o(u|A)dS(u|A,Z) + d\Lambda_0^o(u)e^{\beta^o A}G^o(u|A)S(u|A,Z)\right\}|Z\right]\right),
\end{aligned}
$$

calculating the above conditional expectation considering that $w = \frac{A}{\pi(Z)} + \frac{1-A}{1-\pi(Z)}$, we have:

$$
\begin{aligned}
&\mathrm{E}\left\{U_2^{AIPW}(t,\beta^o,\Lambda_0^o;\pi^o,S,G^o)\right\}\\
&= \int_0^t \mathrm{E}\left\{\frac{\mathrm{E}\{A|Z\}}{\pi^o(Z)}dM^1(u)\right\} + \int_0^t \mathrm{E}\left\{\frac{\mathrm{E}\{1-A|Z\}}{1-\pi^o(Z)}dM^o(u)\right\}\\
&\quad + \sum_{a=0,1}\int_0^t \mathrm{E}\left\{G^o(u|a)dS(u|a,Z) + d\Lambda_0^o(u)e^{\beta^o a}G^o(u|a)S(u|a,Z)\right\}\\
&\quad - \sum_{a=0,1}\int_0^t \mathrm{E}\left\{G^o(u|a)dS(u|a,Z) + d\Lambda_0^o(u)e^{\beta^o a}G^o(u|a)S(u|a,Z)\right\}\\
&= \int_0^t \mathrm{E}\left\{dM^1(u)\right\} + \int_0^t \mathrm{E}\{dM^o(u)\} = 0.
\end{aligned}
$$

On the other hand, if $S = S^o$:

$$
\begin{aligned}
&\mathrm{E}\left\{U_2^{AIPW}(t,\beta^o,\Lambda_0^o;\pi,S^o,G^o)\right\}\\
&= \int_0^t \mathrm{E}\left[\mathrm{E}\{wdM(u)|A,Z\} - \int_0^t \mathrm{E}\{wdM(u)|A,Z\} + \sum_{a=0,1}\int_0^t \mathrm{E}\{dM(u)|A=a,Z\}\right]\\
&= \sum_{a=0,1}\int_0^t \mathrm{E}\{dM^a(u)\} = 0.
\end{aligned}
$$

□

Proof of Theorem 10 and 11

By Taylor expansion of the score $U_{1,n}^{AIPW}\left(\beta;\hat{\pi},\hat{S},\hat{G}\right)$ around β^o we get:

$$
U_{1,n}^{AIPW}\left(\beta;\hat{\pi},\hat{S},\hat{G}\right) = U_{1,n}^{AIPW}(\beta^o;\hat{\pi},\hat{S},\hat{G}) + \frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\hat{\pi},\hat{S},\hat{G})\bigg|_{\beta=\tilde{\beta}}(\beta-\beta^o),
$$

where $\tilde{\beta}$ is a point between β and β^o. Therefore, by construction of $\hat{\beta}$, we have:

$$-\sqrt{n}U_{1,n}^{AIPW}(\beta^o;\hat{\pi},\hat{S},\hat{G}) = +\frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\hat{\pi},\hat{S},\hat{G})\bigg|_{\beta=\tilde{\beta}}\sqrt{n}(\hat{\beta}-\beta^o). \tag{4.26}$$

It can be proved that $\frac{\partial}{\partial\beta}U(\beta;\hat{\pi},\hat{S},\hat{G})$ converges to $\sigma(\beta)$, where:

$$\sigma(\beta) = \int_0^\tau \left\{\left(\frac{s^{(1)}}{s^{(0)}}\right)^2 - \frac{s^{(1)}}{s^{(0)}}\right\}(t;\beta,\pi^*,S^*,G^o)d\Lambda_0^o(t)s^{(0)}(t;\beta^o,\pi^*,S^*,G^o).$$

The next lemma indeed holds.

Lemma 21. *Let model* (4.1) *and Assumptions 2-9 hold, we have:*

$$\frac{\partial}{\partial\beta}U(\beta;\hat{\pi},\hat{S},\hat{G}) = \sigma(\beta)+o_p(1). \tag{4.27}$$

Term $\sqrt{n}U_{1,n}^{AIPW}(\beta^o;\hat{\pi},\hat{S},\hat{G})$ requires a little bit more attention. Indeed, when both models are correctly specified, it converges to $\sqrt{n}U_{1,n}^{AIPW}(\beta^o;\pi^o,S^o,G^o)$. However, when only one of the two models is correctly specified, the limit contains an extra term that depends on which model is correctly specified. The following lemma proves the above in details.

Lemma 22. *Let model* (4.1) *and Assumptions 2-10 hold. If $b_nc_n = o(n^{-1/2})$, it holds:*

$$U_{1,n}^{AIPW}\left(\beta^o;\hat{\pi},\hat{S},\hat{G}\right) \;=\; U_{1,n}^{AIPW}(\beta^o;\pi^*,S^*,G^o)+o_p(1). \tag{4.28}$$

Moreover:

a) If $S^*(t\mid a,z) = S^o(t\mid a,z)$ and $\pi^*(z)=\pi^o(z)$ and $b_nc_n = o(n^{-1/2})$ it holds:

$$U_{1,n}^{AIPW}\left(\beta^o;\hat{\pi},\hat{S},\hat{G}\right) \;=\; U_{1,n}^{AIPW}(\beta^o;\pi^*,S^*,G^o)+o_p(n^{-1/2}). \tag{4.29}$$

b) $\pi^*(z) = \pi^o(z)$, $S^*(t \mid a,z) \neq S^o(t \mid a,z)$ with $b_n = n^{-1/2}$; specifically there exists an influence function $\phi(z)$ such that $\hat{\pi}(z) - \pi^*(z) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(z)$, we have:

$$U_{1,n}^{AIPW}\left(\beta^o;\hat{\pi},\hat{S},\hat{G}\right) \tag{4.30}$$

$$\begin{aligned}
= \; & U_{1,n}^{AIPW}\left(\beta^o;\pi^*,S^*,G\right) + o_p(n^{-1/2}) \\
& -\int_0^\tau \frac{1}{n^2}\sum_{i,j=1}^{n}\phi_j(Z_i)\left\{\pi_i^*\right\}^{-2}A_i\left(dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)e^{\beta^o A_i}R_i(t,S^*,G^o)\right) \\
& -\int_0^\tau \left\{s^{(0)}(t;\beta^o,\hat{\pi},S^*,G^o)\right\}^{-1}\left\{s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)\right. \\
& \quad \left. -\frac{1}{n^2}\sum_{l,m=1}^{n}\phi_l(Z_m)\left\{\pi_m^*\right\}^{-2}A_m e^{\beta^o A_m}R_m(t,S^*,G^o)\right\} \\
& \quad \times\frac{1}{n^2}\sum_{i,j=1}^{n}\phi_j(Z_i)\left[\frac{1-A_i}{\left\{1-\pi_i^*\right\}^2} - \frac{A_i}{\left\{\pi_i^*\right\}^2}\right] \\
& \quad \times\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)\exp(\beta^o A_i)R_i(t,S^*,G^o)\right\}.
\end{aligned} \tag{4.31}$$

c) $S^*(t|a,z) = S^o(t|a,z)$, $\pi^*(z) \neq \pi^o(z)$ and $c_n = n^{-1/2}$; specifically there exists an influence *function* $\psi(t,a,z)$ *such that* $\hat{S}(t|a,z)-S^{*}(t|a,z)=\frac{1}{n}\sum_{i=1}^{n}\psi_i(t,a,z)$*, we have:*

$$\begin{aligned}
& U_{1,n}^{AIPW}\left(\beta^{o};\hat{\pi},\hat{S},\hat{G}\right) \qquad\qquad (4.32)\\
= \; & U_{1,n}^{AIPW}\left(\beta^{o};\pi^{*},S^{*},G\right)+o_p(n^{-1/2})\\
& -\int_0^{\tau}\left\{s^{(0)}(t;\beta^{o},\pi^{*},S^{*},G^{o})\right\}^{-1}\left[s^{(1)}(t;\beta^{o},\pi^{*},S^{*},G^{o})+\frac{1}{n}\sum_{l=1}^{n}J_l(t,1)\left\{\frac{A_l}{\pi_l^{*}}e^{\beta^{o}A_l}-e^{\beta^{o}}\right\}\right]\\
& \times\frac{1}{n^2}\sum_{i,j=1}^{n}\Big(w_i^{*}\left\{G^{o}(t,A_i)d\psi_j(t,A_i,Z_i)+dS_i^{o}(t,A_i)\rho_j(t,A_i)\right\}\\
& -\sum_{a=0,1}\left\{G^{o}(t,a)d\psi_j(t,a,Z_i)+dS_i^{o}(t,a)\rho_j(t,a)\right\}\\
& +d\Lambda_0^{o}(u)\Big\{w_i^{*}e^{\beta^{o}A_i}\left\{G^{o}(t,A_i)\psi_j(t,A_i,Z_i)+S_i^{o}(t,A_i)\rho_j(t,A_i)\right\}\\
& -\sum_{a=0,1}e^{\beta^{o}a}\left\{G^{o}(t,a)\psi_j(t,a,Z_i)+S_i^{o}(t,a)\rho_j(t,a)\right\}\Bigg)\\
& +\int_0^{\tau}d\Lambda_0^{o}(t)\frac{1}{n^2}\sum_{i,j=1}^{n}\left\{G^{o}(t,1)\psi_j(t,1,Z_i)+S_i^{o}(t,1)\rho_j(t,1)\right\}\left\{\frac{A_i}{\pi_i^{*}}e^{\beta^{o}A_i}-e^{\beta^{o}}\right\}\\
& +\int_0^{\tau}\frac{1}{n^2}\sum_{i,j=1}^{n}\left\{G^{o}(t,A_i)d\psi_j(t,1,Z_i)+dS_i^{o}(t,A_i)\rho_j(t,1)\right\}\left\{\frac{A_i}{\pi_i^{*}}-1\right\}.
\end{aligned}$$

Putting together (4.26) and the results of the previous lemma we will prove $\sqrt{n}(\hat{\beta}-\beta^{o})$ can be written as a sum of i.i.d mean zero terms and so the consistency and the asymptotic normality of $\hat{\beta}$ follows. We now report the details of the proofs of Theorem 10 and 11. Proof of Lemma 21 and 22 are reported in Section 4.8.3.

Proof of Theorem 10. By Taylor expansion and by Lemma 21, 22, for β in a neighborhood of β^{o}

we have:

$$U_{1,n}^{AIPW}\left(\beta;\hat{\pi},\hat{S},\hat{G}\right) = U_{1,n}^{AIPW}(\beta^o;\hat{\pi},\hat{S},\hat{G}) + \frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\hat{\pi},\hat{S},\hat{G})\bigg|_{\beta=\tilde{\beta}}(\beta-\beta^o) \tag{4.33}$$

$$= U_{1,n}^{AIPW}(\beta^o;\hat{\pi},\hat{S},\hat{G}) + \sigma(\tilde{\beta})(\beta-\beta^o) + o_p(1), \tag{4.34}$$

where $\tilde{\beta}$ is a point between β and β^o.

We notice that, by considering a finite τ and by Assumptions 8 and 9, we have

$$\left|\sigma(\tilde{\beta})\right| > 0. \tag{4.35}$$

By double robustness of the score, (Theorem 9) and, by application of Hoeffding's inequality, we have:

$$U_{1,n}^{AIPW}(\beta^o;\pi^*,S^*,G^o) = O_p(n^{-1/2}). \tag{4.36}$$

if either $\pi^*(\cdot) = \pi^o(\cdot)$ or $S^*(\cdot) = S^o(\cdot)$. Therefore, putting together (4.33) and (4.36) we have for any $|\delta| < \frac{1}{2}$:

$$\begin{aligned} U_{1,n}^{AIPW}\left(\beta^o \pm n^{-\delta};\hat{\pi},\hat{S},\hat{G}\right) &= U_{1,n}^{AIPW}(\beta^o;\pi^*,S^*,G^o) \pm n^{-\delta}\sigma(\tilde{\beta}) + o_p(1) \\ &= \pm n^{-\delta}\sigma(\tilde{\beta}) + O_p(n^{-1/2}). \end{aligned}$$

Therefore, by the above and by (4.35), we have:

$$U_{1,n}^{AIPW}\left(\beta^o - n^{-\delta};\hat{\pi},\hat{S},\hat{G}\right) < 0 < U_{1,n}^{AIPW}\left(\beta^o + n^{-\delta};\hat{\pi},\hat{S},\hat{G}\right),$$

or

$$U_{1,n}^{AIPW}\left(\beta^o + n^{-\delta}; \hat{\pi}, \hat{S}, \hat{G}\right) < 0 < U_{1,n}^{AIPW}\left(\beta^o - n^{-\delta}; \hat{\pi}, \hat{S}, \hat{G}\right).$$

Hence, by construction of $\hat{\beta}$, we can conclude that $\hat{\beta} - \beta^o = O_p(n^{-\delta}) = o_p(1)$. □

Proof of Theorem 11. By Taylor expansion, Lemma 21 and consistency of $\hat{\beta}$ we get:

$$-\sqrt{n}U_{1,n}^{AIPW}(\beta^o; \hat{\pi}, \hat{S}, \hat{G}) = \sqrt{n}(\hat{\beta} - \beta^o)\sigma(\beta^o) + o_p(1).$$

We are now left with working on term $U_{1,n}^{AIPW}(\beta^o; \hat{\pi}, \hat{S}, \hat{G})$. We will prove that it can be written as a sum of i.i.d mean zero terms. We divide the proof in the three different scenarios of the Theorem according to which model is correct.

- a) $\pi^* = \pi^o$ and $S^* = S^o$ and $b_n c_n = o(n^{-1/2})$.

  By Lemma 22 we have:

$$\sqrt{n}(\hat{\beta} - \beta^o) = \frac{\sqrt{n}U_{1,n}^{AIPW}(\beta^o; \pi^o, S^o, G^o)}{\sigma(\beta^o)} + o_p(1). \tag{4.37}$$

We focus now on term $U_{1,n}^{AIPW}(\beta^o; \pi^o, S^o, G^o)$. We remind the reader that $-G^o(t|A,Z)dS^o(t|A,Z)- = \mathrm{E}[dN(t)|A,Z]$ and therefore we have:

$$\begin{aligned} &U_{1,n}^{AIPW}(\beta^o; \pi^o, S^o, G^o) \\ &\quad = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big[w_i^o \left\{A_i - \bar{A}(t; \beta^o, \pi^o, S^o, G^o)\right\}\left\{dN_i(t) - \mathrm{E}[dN_i(t)|A_i, Z_i]\right\} \\ &\qquad + \sum_{a=0,1}\left\{a - \bar{A}(t; \beta^o, \pi^o, S^o, G^o)\right\}\mathrm{E}[dN_i(t)|a, Z_i]\Bigg]. \end{aligned}$$

Moreover, noticing that, by definition of $\bar{A}$ and algebra, we get:

$$\begin{aligned}\int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big[& w_i^o e^{\beta^o A_i}\left\{A_i - \bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\}\{Y_i(t) - \mathrm{E}[Y_i(t)|A_i,Z_i]\} \\ & + \sum_{a=0,1} e^{\beta^o a}\left\{a - \bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\}\mathrm{E}[Y_i(t)|a,Z_i]\Bigg] = 0,\end{aligned}$$

we can conclude that:

$$\begin{aligned}& U_{1,n}^{AIPW}(\beta^o;\pi^o,S^o,G^o) \\ & = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \big(w_i^o\left\{A_i - \bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\}[dM_i(t) - \mathrm{E}\{dM_i(t)|A_i,Z_i\}] \\ & \quad + \sum_{a=0,1}\left\{a - \bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\}\mathrm{E}\{dM_i(t)|A=a,Z_i\}\Bigg) \\ & = Q_1 + Q_2,\end{aligned}$$

where

$$\begin{aligned}Q_1 = \int_0^\tau \frac{1}{n}\sum_{i=1}^n & (w_i^o\{A_i - \bar{a}(t;\beta^o,\pi^o,S^o,G^o)\}[dM_i(t) - \mathrm{E}\{dM_i(t)|A_i,Z_i\}] \\ & + \sum_{a=0,1}\{a - \bar{a}(t;\beta^o,\pi^o,S^o,G^o)\}\mathrm{E}\{dM_i(t)|a,Z_i\}\Bigg),\end{aligned}$$

$$\begin{aligned}Q_2 = \int_0^\tau & \left\{a(t;\beta^o,\pi^o,S^o,G^o) - \bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\} \\ & \times \frac{1}{n}\sum_{i=1}^n\left[w_i^o dM_i(t) - w_i^o\mathrm{E}\{dM_i(t)|A_i,Z_i\} + \sum_{a=0,1}\mathrm{E}\{dM_i(t)|a,Z_i\}\right].\end{aligned}$$

Q_1 is the leading term since it is a sum of i.i.d mean zero terms. $Q_2 = o_p(n^{-1/2})$ by Assumption 9 and by the fact that $\mathrm{E}\left[w^o M(t) - w^o \mathrm{E}\{M(t)|A,Z\} + \sum_{a=0,1} \mathrm{E}\{M(t)|A=a,Z\}\right] = 0$. Therefore we have:

$$\sqrt{n}(\hat{\beta} - \beta^o) = \sigma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_i^a + o_p(1), \tag{4.38}$$

where

$$\begin{aligned} \varphi_i^a = & \int_0^\tau w_i^o \{A_i - \bar{a}(t;\beta^o,\pi^o,S^o,G^o)\} [dM_i(t) - \mathrm{E}\{dM_i(t)|A_i,Z_i\}] \\ & + \sum_{a=0,1} \{a - \bar{a}(t;\beta^o,\pi^o,S^o,G^o)\} \mathrm{E}\{dM_i(t)|a,Z_i\}, \end{aligned}$$

and $\sigma = \sigma(\beta^o)$.

- b) $\pi^*(z) = \pi^o(z)$, $S^*(t|a,z) \neq S^o(t|a,z)$ and $b_n = n^{-1/2}$; specifically there exists an influence function $\phi(z)$ such that $\hat{\pi}(z) - \pi^*(z) = \frac{1}{n}\sum_{i=1}^n \phi_i(z) + o_p(n^{-1/2})$.

  Similarly to part a) we have:

$$\begin{aligned} & U_{1,n}^{AIPW}(\beta^o;\pi^o,S^*,G^o) \\ & = \int_0^\tau \frac{1}{n} \sum_{i=1}^{n} [w_i^o \{A_i - \bar{a}(t;\beta^o,\pi^o,S^*,G^o)\} \\ & \quad \times \left\{dM_i(t) + dS_i^*(t,A_i)G^o(t,A_i) + e^{\beta^o A_i} S_i^*(t,A_i)G^o(t,A_i)dt\right\} \\ & \quad - \sum_{a=0,1} \left\{a - \bar{a}(t;\beta^o,\pi^o,\hat{S},G^o)\right\} \left\{dS_i^*(t,a)G^o(t,a) + e^{\beta^o a} S_i^*(t,a)G^o(t,a)dt\right\}\Bigg] \\ & \quad + o_p(n^{-1/2}). \end{aligned}$$

Therefore, by the above and by Lemma 22, we have:

$$\sqrt{n}(\hat{\beta}-\beta^o)=\sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i^b+o_p(1), \tag{4.39}$$

where

$$\begin{aligned}
\varphi_i^b=&\int_0^\tau w_i^o\left\{A_i-\bar{a}(t;\beta^o,\pi^o,S^*,G^o)\right\}\left\{dM_i(t)+dS_i^*(t,A_i)G^o(t,A_i)+e^{\beta^o A_i}S_i^*(t,A_i)G^o(t,A_i)dt\right\}\\
&-\sum_{a=0,1}\left\{a-\bar{a}(t;\beta^o,\pi^o,S^*,G^o)\right\}\left\{dS_i^*(t,a)G^o(t,a)+e^{\beta^o a}S_i^*(t,a)G^o(t,a)dt\right\}\\
&-\int_0^\tau\frac{1}{n}\sum_{j=1}^{n}\phi_i(Z_j)\left\{\pi_j^*\right\}^{-2}A_j\left(dN_j(t)+G^o(t,A_j)dS_j^*(t,A_j)-d\Lambda_0^o(t)e^{\beta^o A_j}R_j(t,S^*,G^o)\right)\\
&-\int_0^\tau\left\{s^{(0)}(t;\beta^o,\hat{\pi},S^*,G^o)\right\}^{-1}\left\{s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)\right.\\
&\left.-\frac{1}{n^2}\sum_{l,m=1}^{n}\phi_l(Z_m)\left\{\pi_m^*\right\}^{-2}A_m e^{\beta^o A_m}R_m(t,S^*,G^o)\right\}\\
&\times\frac{1}{n}\sum_{j=1}^{n}\phi_i(Z_j)\left[\frac{1-A_j}{\left\{1-\pi_j^*\right\}^2}-\frac{A_j}{\left\{\pi_j^*\right\}^2}\right]\\
&\times\left\{dN_j(t)+G^o(t,A_j)dS_j^*(t,A_j)-d\Lambda_0^o(t)\exp(\beta^o A_j)R_j(t,S^*,G^o)\right\}.
\end{aligned}$$

- c) $S^*(t\mid a,z)=S^o(t\mid a,z)$, $\pi^*(z)\neq\pi^o(z)$ and $c_n=n^{-1/2}$, specifically, there exists an influence function $\psi(t,a,z)$ such that $\hat{S}(t\mid a,z)-S^*(t\mid a,z)=\frac{1}{n}\sum_{i=1}^{n}\psi_i(t,a,z)+o_p(n^{-1/2})$.

By Lemma 22, we have: Similarly to part a) we have:

$$
\begin{aligned}
& U_{1,n}^{AIPW}(\beta^o;\pi^*,S^o,G^o) \\
& \quad = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Big[w_i^* \{A_i - \bar{a}(t;\beta^o,\pi^*,S^o,G^o)\} \\
& \qquad \times \Big\{ dM_i(t) + dS_i^o(t,A_i)G^o(t,A_i) + e^{\beta^o A_i} S_i^o(t,A_i)G^o(t,A_i)dt \Big\} \\
& \qquad - \sum_{a=0,1} \{a - \bar{a}(t;\beta^o,\pi^*,S^o,G^o)\} \Big\{ dS_i^*(t,a)G^o(t,a) + e^{\beta^o a} S_i^o(t,a)G^o(t,a)dt \Big\} \Big] \\
& \qquad + o_p(n^{-1/2}).
\end{aligned}
$$

Therefore, by the above and by Lemma 22, we have:

$$
\sqrt{n}(\hat{\beta} - \beta^o) = \sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_i^c + o_p(1), \tag{4.40}
$$

where

$$
\begin{aligned}
\varphi_i^c = & \int_0^\tau w_i^* \left\{A_i - \bar{a}(t;\beta^o,\pi^*,S^o,G^o)\right\} \\
& \times \left\{ dM_i(t) + dS_i^o(t,A_i)G^o(t,A_i) + e^{\beta^o A_i} S_i^o(t,A_i)G^o(t,A_i)d\Lambda_0^o(t) \right\} \\
& - \int_0^\tau \sum_{a=0,1} \left\{a - \bar{a}(t;\beta^o,\pi^*,S^o,G^o)\right\} \left\{ dS_i^o(t,a)G^o(t,a) + e^{\beta^o a} S_i^o(t,a)G^o(t,a)d\Lambda_0^o(t) \right\} \\
& - \int_0^\tau \left\{ s^{(0)}(t;\beta^o,\pi^*,S^o,G^o) \right\}^{-1} \\
& \times \left[s^{(1)}(t;\beta^o,\pi^*,S^o,G^o) + \frac{1}{n}\sum_{l=1}^n J_l(t,1) \left\{ \frac{A_l}{\pi_l^*} e^{\beta^o A_l} - e^{\beta^o} \right\} \right] \\
& \times \frac{1}{n}\sum_{j=1}^n \Big(w_j^* \left\{ G^o(t,A_j)d\psi_i(t,A_j,Z_j) + dS_j^o(t,A_j)\rho_i(t,A_j) \right\} \\
& - \sum_{a=0,1} \left\{ G^o(t,a)d\psi_i(t,a,Z_j) + dS_j^o(t,a)\rho_i(t,a) \right\} \\
& + d\Lambda_0^o(u) \Big[w_j^* e^{\beta^o A_j} \left\{ G^o(t,A_j)\psi_i(t,A_j,Z_j) + S_j^o(t,A_j)\rho_i(t,A_j) \right\} \\
& - \sum_{a=0,1} e^{\beta^o a} \left\{ G^o(t,a)\psi_i(t,a,Z_j) + S_j^o(t,a)\rho_i(t,a) \right\} \Big] \Big) \\
& + \int_0^\tau d\Lambda_0^o(t) \frac{1}{n}\sum_{j=1}^n \left\{ G^o(t,1)\psi_i(t,1,Z_j) + S_j^o(t,1)\rho_i(t,1) \right\} \left\{ \frac{A_i}{\pi_i^*} e^{\beta^o A_i} - e^{\beta^o} \right\} \\
& + \int_0^\tau \frac{1}{n}\sum_{j=1}^n \left\{ G^o(t,A_j)d\psi_i(t,1,Z_j) + dS_j^o(t,A_j)\rho_i(t,1) \right\} \left\{ \frac{A_j}{\pi_j^*} - 1 \right\}.
\end{aligned}
$$

Putting together both (4.38), (4.39) and (4.40) we have the more general (4.17). Therefore, $\sqrt{n}(\hat{\beta} - \beta^o)$ is a sum of i.i.d. mean zero terms and we can apply Multivariate Central Limit Theorem to prove its asymptotic normality. □

Proof of Theorem 12

Proof of Theorem 12. We will prove separately that $\hat{V}$ and $\hat{\sigma}$ are consistent estimators for $Var(\varphi)$ and σ respectively.

a) We remind the reader that when both models are correct $\varphi = \varphi^a(\beta^o, \pi^o, S^o, G^o)$. We have:

$$\left|\hat{V} - Var(\varphi)\right| = \left|\hat{V} - E\left\{\varphi^a(\beta^o, \pi^o, S^o, G^o)\right\}^2\right| \leq Q_1 + Q_2 + Q_3,$$

where

$$Q_1 = \left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\hat{\varphi}_i^a(\hat{\beta}, \hat{\pi}, \hat{S}, \hat{G})\right\}^2 - \left\{\hat{\varphi}_i^a(\beta^o, \pi^o, S^o, G^o)\right\}^2\right]\right|,$$

$$Q_2 = \left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\hat{\varphi}_i^a(\beta^o, \pi^o, S^o, G^o)\right\}^2 - \left\{\varphi_i^a(\beta^o, \pi^o, S^o, G^o)\right\}^2\right]\right|,$$

$$Q_3 = \left|\frac{1}{n}\sum_{i=1}^{n}\left\{\varphi_i^a(\beta^o, \pi^o, S^o, G^o)\right\}^2 - E\left\{\varphi^a(\beta^o, \pi^o, S^o, G^o)\right\}^2\right|.$$

By Assumptions 7 and 8 and continuos mapping Theorem, $Q_1 = o_p(1)$ follows.
By continuos mapping Theorem and Assumptions 8 and 9 $Q_2 = o_p(1)$.
By construction and by Assumption 8, by law of large numbers it is easy to see that $Q_3 = o_p(1)$.
Therefore $\hat{V} = Var(\varphi) + o_p(1)$.

b) We have:

$$|\hat{\sigma} - \sigma| \leq Q_1 + Q_2 + Q_3,$$

where

$$\begin{aligned} Q_1 = \frac{1}{n}\sum_{i=1}^{n}\Bigg| & \int_0^{\tau}\left[\left\{\bar{A}(t;\hat{\beta},\hat{\pi},\hat{S},\hat{G})\right\}^2 - \bar{A}(t;\hat{\beta},\hat{\pi},\hat{S},\hat{G})\right] \\ & \quad \times \left[\hat{w}_i dN_i(t) + \hat{w}_i \hat{G}(t,A_i) d\hat{S}_i(t,A_i) - \sum_{a=0,1}\hat{G}(t,a) d\hat{S}_i(t,a)\right] \\ & - \int_0^{\tau}\left[\left\{\bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\}^2 - \bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right] \\ & \quad \times \left[w_i^o dN_i(t) + w_i^o G^o(t,A_i) dS_i^o(t,A_i) - \sum_{a=0,1} G^o(t,a) dS_i^o(t,a)\right]\Bigg|, \end{aligned}$$

and

$$\begin{aligned} Q_2 = \frac{1}{n}\sum_{i=1}^{n}\Bigg| \int_0^{\tau} & \Big[\left\{\bar{A}(t;\beta^o,\pi^o,S^o,G^o)\right\}^2 - \left\{\bar{a}(t;\beta^o,\pi^o,S^o,G^o)\right\}^2 \\ & \quad -\bar{A}(t;\beta^o,\pi^o,S^o,G^o) + \bar{a}(t;\beta^o,\pi^o,S^o,G^o)\Big] \\ & \times \left[w_i^o dN_i(t) + w_i^o G^o(t,A_i) dS_i^o(t,A_i) - \sum_{a=0,1} G^o(t,a) dS_i^o(t,a)\right]\Bigg|, \end{aligned}$$

and

$$\begin{aligned} Q_3 = \Bigg| \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} & \left[\left\{\bar{a}(t;\beta^o,\pi^o,S^o,G^o)\right\}^2 - \bar{a}(t;\beta^o,\pi^o,S^o,G^o)\right] \\ & \times \Bigg[w_i^o dN_i(t) + w_i^o G^o(t,A_i) dS_i^o(t,A_i) - \sum_{a=0,1} G^o(t,a) dS_i^o(t,a) \\ & \quad - d\Lambda_0^o(t) s^{(0)}(t;\beta^o,\pi^o,S^o,G^o)\Bigg]\Bigg|. \end{aligned}$$

Again by Assumptions 7 and 8 and continuos mapping Theorem, $Q_1 = o_p(1)$ follows.

By continuos mapping Theorem and Assumptions 8 and 9 $Q_2 = o_p(1)$.

For the last term we have:

$$
\begin{aligned}
&w_i^o dN_i(t) + w_i^o G^o(t, A_i) dS_i^o(t, A_i) - \sum_{a=0,1} G^o(t,a) dS_i^o(t,a) - d\Lambda_0^o(t) s^{(0)}(t; \beta^o, \pi^o, S^o, G^o) \\
&\quad = w_i^o dM_i(t) - w_i^o \mathrm{E}\{dM(t)|A_i, Z_i\} + \sum_{a=0,1} \mathrm{E}\{dM(t)|A_i, Z_i\},
\end{aligned}
$$

and therefore the above has mean zero. By an application of Bernstein's inequality to the bounded random variable:

$$
\begin{aligned}
&\int_0^\tau \left[\{\bar{a}(t; \beta^o, \pi^o, S^o, G^o)\}^2 - \bar{a}(t; \beta^o, \pi^o, S^o, G^o) \right] \\
&\quad \times \left[w_i^o dM_i(t) - w_i^o \mathrm{E}\{dM(t)|A_i, Z_i\} + \sum_{a=0,1} \mathrm{E}\{dM(t)|A_i, Z_i\} \right],
\end{aligned}
$$

we can therefore conclude that $Q_3 = o_p(1)$. □

4.8.3 Proof of Lemmas

Proof of Lemma 21

Proof of Lemma 21. Simple algebra gives us:

$$
\begin{aligned}
\frac{\partial}{\partial \beta} U(\beta; \pi, S, G) = &\int_0^\tau \frac{1}{n} \sum_{i=1}^n \left[\left\{\bar{A}(t; \beta, \pi, S, G)\right\}^2 - \bar{A}(t; \beta, \pi, S, G) \right] \\
&\times \left[w_i \{dN_i(t) + G(t, A_i) dS_i(t, A_i)\} - \sum_{a=0,1} G(t,a) dS_i(t,a) \right].
\end{aligned}
$$

First, we prove that $\frac{\partial}{\partial\beta}U(\beta;\hat{\pi},\hat{S},\hat{G}) = \frac{\partial}{\partial\beta}U(\beta;\pi^*,S^*,G^o) + o_p(1)$.

We consider the following decomposition:

$$\begin{aligned}
\frac{\partial}{\partial\beta}U(\beta;\hat{\pi},\hat{S},\hat{G}) &= \frac{\partial}{\partial\beta}U(\beta;\pi^*,S^*,G^o) \\
&\quad + \frac{\partial}{\partial\beta}U\left(\beta;\hat{\pi},\hat{S},\hat{G}\right) - \frac{\partial}{\partial\beta}U\left(\beta;\pi^*,\hat{S},\hat{G}\right) \\
&\quad + \frac{\partial}{\partial\beta}U\left(\beta;\pi^*,\hat{S},\hat{G}\right) - \frac{\partial}{\partial\beta}U\left(\beta;\pi^*,S^*,G^o\right) \\
&= Q_1 + Q_2 + Q_3.
\end{aligned}$$

We now prove separately that $Q_2 = o_p(1)$ and $Q_3 = o_p(1)$. From now on, for ease of notation, we use:

$$\left\{(\bar{A})^2 - \bar{A}\right\}(t;\beta,\pi,S,G) = \left\{\bar{A}(t;\beta,\pi,S,G)\right\}^2 - \bar{A}(t;\beta,\pi,S,G).$$

- Term Q_2:

By algebra, we get:

$$\begin{aligned}
Q_2 = \int_0^\tau \frac{1}{n}\sum_{i=1}^{n} &\left[\hat{w}_i\left\{(\bar{A})^2 - \bar{A}\right\}(t;\beta,\hat{\pi},\hat{S},\hat{G}) - w_i^*\left\{(\bar{A})^2 - \bar{A}\right\}(t;\beta,\pi^*,\hat{S},\hat{G})\right] \\
&\times\left\{dN_i(t) + \hat{G}(t,A_i)d\hat{S}_i(t,A_i)\right\}.
\end{aligned}$$

By adding and subtracting S^o and G^o:

$$
\begin{aligned}
Q_2 = & \int_0^\tau \frac{1}{n} \sum_{i=1}^n \left[\hat{w}_i \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\hat{\pi},\hat{S},\hat{G}) - w_i^* \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\pi^*,\hat{S},\hat{G}) \right] \\
& \times \left\{ dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i) \right\} \\
& + \int_0^\tau \frac{1}{n} \sum_{i=1}^n \left[\hat{w}_i \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\hat{\pi},\hat{S},\hat{G}) - w_i^* \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\pi^*,\hat{S},\hat{G}) \right] \\
& \times \left\{ \hat{G}(t,A_i) d\hat{S}_i(t,A_i) - G^o(t,A_i) dS_i^*(t,A_i) \right\} \\
= & Q_{21} + Q_{22},
\end{aligned}
$$

where

$$
\begin{aligned}
Q_{21} = & \int_0^\tau \frac{1}{n} \sum_{i=1}^n \left(\hat{w}_i \left[\left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\hat{\pi},\hat{S},\hat{G}) - \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\pi^*,\hat{S},\hat{G}) \right] \right. \\
& \left. + (\hat{w}_i - w_i^*) \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\pi^*,\hat{S},\hat{G}) \right] \left\{ dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i) \right\},
\end{aligned}
$$

and

$$
\begin{aligned}
Q_{22} = + & \int_0^\tau \frac{1}{n} \sum_{i=1}^n \left(\hat{w}_i \left[\left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\hat{\pi},\hat{S},\hat{G}) - \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\pi^*,\hat{S},\hat{G}) \right] \right. \\
& \left. + (\hat{w}_i - w_i^*) \left\{ (\bar{A})^2 - \bar{A} \right\} (t;\beta,\pi^*,\hat{S},\hat{G}) \right) \left\{ \hat{G}(t,A_i) d\hat{S}_i(t,A_i) - G^o(t,A_i) dS_i^*(t,A_i) \right\}.
\end{aligned}
$$

We notice how, $\bar{A}(t;\beta,\pi^*,\hat{S},\hat{G}) = \frac{s^{(1)}(t;\beta)}{s^{(0)}(t;\beta)}$ and by the fact that $\tau < \infty$ and by Assumption 8, everything is bounded. Therefore, by Assumption 7, it is easy to see that $\bar{A}(t;\beta,\hat{\pi},\hat{S},\hat{G}) - \bar{A}(t;\beta,\pi^*,\hat{S},\hat{G}) = o_p(1)$. By this and by Assumption 7, we have $Q_{21} = o_p(1)$ and $Q_{22} = o_p(1)$ and so $Q_2 = o_p(1)$.

- Term Q_3:

We have:

$$
\begin{aligned}
Q_3 = \int_0^\tau \frac{1}{n}\sum_{i=1}^{n} w_i^* \big[& \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,\hat{S},\hat{G})\big\{dN_i(t) + \hat{G}(t,A_i)d\hat{S}_i(t,A_i)\big\} \\
& \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,S^*,G^o)\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i)\}\big] \\
& - \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1}\int_0^\tau \big[\big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,\hat{S},\hat{G})\hat{G}(t,a)d\hat{S}_i(t,a) \\
& - \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,S^*,G^o)G^o(t,a)dS_i^*(t,a)\big].
\end{aligned}
$$

By algebra we have:

$$
\begin{aligned}
Q_3 = & \int_0^\tau \frac{1}{n}\sum_{i=1}^{n} \big[\big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,\hat{S},\hat{G}) - \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,S^*,G^o)\big] \\
& \times \left[w_i^* \{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i)\} - \sum_{a=0,1} G^o(t,a)dS_i^*(t,a) \right] \\
& + \int_0^\tau \frac{1}{n}\sum_{i=1}^{n} w_i^* \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,\hat{S},\hat{G})\big\{\hat{G}(t,A_i)d\hat{S}_i(t,A_i) - G^o(t,A_i)dS_i^*(t,A_i)\big\} \\
& - \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1}\int_0^\tau \big\{\big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,\hat{S},\hat{G}) - \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,S^*,G^o)\big\} \\
& \quad \times G^o(t,a)dS_i^*(t,a) \\
& - \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1}\int_0^\tau \big\{(\bar{A})^2 - \bar{A}\big\}(t;\beta,\pi^*,\hat{S},\hat{G})\big\{\hat{G}(t,a)d\hat{S}_i(t,a) - G^o(t,a)dS_i^*(t,a)\big\} \\
= {} & Q_{31} + Q_{32} + Q_{33} + Q_{34},
\end{aligned}
$$

where

$$Q_{31} = \int_0^\tau \left[\left\{ (\bar{A})^2 - \bar{A} \right\} (t; \beta, \pi^*, \hat{S}, \hat{G}) - \left\{ (\bar{A})^2 - \bar{A} \right\} (t; \beta, \pi^*, S^*, G^o) \right] \\ \times \frac{1}{n} \sum_{i=1}^{n} \left[w_i^* \left\{ dN_i(t) + G^o(t, A_i) dS_i^*(t, A_i) \right\} - \sum_{a=0,1} G^o(t, a) dS_i^*(t, a) \right],$$

$$Q_{32} = \int_0^\tau \frac{1}{n} \sum_{i=1}^{n} w_i^* \left\{ (\bar{A})^2 - \bar{A} \right\} (t; \beta, \pi^*, S^*, G^o) \left\{ \hat{G}(t, A_i) d\hat{S}_i(t, A_i) - G^o(t, A_i) dS_i^*(t, A_i) \right\},$$

$$Q_{33} = \int_0^\tau \left[\left\{ (\bar{A})^2 - \bar{A} \right\} (t; \beta, \pi^*, \hat{S}, \hat{G}) - \left\{ (\bar{A})^2 - \bar{A} \right\} (t; \beta, \pi^*, S^*, G^o) \right] \\ \times \frac{1}{n} \sum_{i=1}^{n} \left[w_i^* \left\{ \hat{G}(t, A_i) d\hat{S}_i(t, A_i) - G^o(t, A_i) dS_i^*(t, A_i) \right\} \right. \\ \left. - \sum_{a=0,1} \left\{ \hat{G}(t, a) d\hat{S}_i(t, a) - G^o(t, a) dS_i^*(t, a) \right\} \right],$$

$$Q_{34} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{a=0,1} \int_0^\tau \left\{ (\bar{A})^2 - \bar{A} \right\} (t; \beta, \pi^*, S^*, G^o) \left\{ \hat{G}(t, a) d\hat{S}_i(t, a) - G^o(t, a) dS_i^*(t, a) \right\}.$$

Similarly to before, by Assumptions 7 and 8, we have $\bar{A}(t; \beta, \pi^*, \hat{S}, \hat{G}) - \bar{A}(t; \beta, \pi^*, S^*, G^*) = o_p(1)$.

Therefore $Q_{31} = o_p(1)$, $Q_{32} = o_p(1)$, $Q_{33} = o_p(1)$ and $Q_{34} = o_p(1)$ and so $Q_3 = o_p(1)$.

We therefore have proved that

$$\frac{\partial}{\partial \beta} U_{1,n}^{AIPW} \left(\beta; \hat{\pi}, \hat{S}, \hat{G} \right) = \frac{\partial}{\partial \beta} U_{1,n}^{AIPW} \left(\beta; \pi^*, S^*, G^o \right) + o_p(1). \tag{4.41}$$

We are left to prove that $\frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\pi^*,S^*,G^o)=\sigma(\beta)+o_p(1)$.

By Assumption 9, we have:

$$\begin{aligned}\frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\pi^*,S^*,G^o) &= \int_0^\tau \frac{1}{n}\sum_{i=1}^{n}\left[\{\bar{a}(t;\beta,\pi^*,S^*,G^o)\}^2-\bar{a}(t;\beta,\pi^*,S^*,G^o)\right]\\ &\quad\times\left[w_i^*\{dN_i(t)+G^o(t,A_i)dS_i^*(t,A_i)\}-\sum_{a=0,1}G^o(t,a)dS_i^*(t,a)\right]\\ &\quad+o_p(1).\end{aligned}$$

Algebra gives us the following:

$$\begin{aligned}&\frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\pi^*,S^*,G^o)\\ &=\int_0^\tau \frac{1}{n}\sum_{i=1}^{n}\left[\{\bar{a}(t;\beta,\pi^*,S^*,G^o)\}^2-\bar{a}(t;\beta,\pi^*,S^*,G^o)\right]\\ &\quad\times\left[w_i^o\left\{dM_i(t)+G^o(t,A_i)dS_i^*(t,A_i)+d\Lambda_0^o(t)e^{(\beta^oA_i)}G^o(t,A_i)S_i^*(t,A_i)\right\}\right.\\ &\qquad\left.-\sum_{a=0,1}\left\{G^o(t,a)dS_i^*(t,a)+d\Lambda_0^o(t)e^{\beta^oa}G^o(t,a)S_i^*(t,a)\right\}+d\Lambda_0^o(t)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right]\\ &\quad+o_p(1).\end{aligned}$$

By double robustness, if either $\pi^*=\pi^o$ or $S^*=S^o$ we have

$$\begin{aligned}&\mathrm{E}\left[w_i^*\left\{dM_i(t)+G^o(t,A_i)dS_i^*(t,A_i)+d\Lambda_0^o(t)e^{(\beta^oA_i)}G^o(t,A_i)S_i^*(t,A_i)\right\}\right.\\ &\quad\left.-\sum_{a=0,1}\left\{G^o(t,a)dS_i^*(t,a)+d\Lambda_0^o(t)e^{\beta^oa}G^o(t,a)S_i^*(t,a)\right\}\right]=0,\end{aligned}$$

and therefore, by the above and by Assumption 9, we have

$$\frac{\partial}{\partial\beta}U_{1,n}^{AIPW}(\beta;\pi^*,S^*,G^o)=\sigma(\beta)+o_p(1). \tag{4.42}$$

By (4.41) and (4.42), the lemma is proved. □

Proof of Lemma 22

The proof of Lemma 22 requires the following additional lemma:

Lemma 23. *Under Assumption 9 we have:*

$$\begin{aligned}
&\tilde\Lambda_0(t,\beta^o,\pi^*,S^*,G^o)-\Lambda_0^o(t) \qquad (4.43)\\
&=\int_0^t\frac{1}{n}\sum_{i=1}^n\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\quad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}.
\end{aligned}$$

$$\begin{aligned}
&\tilde\Lambda_0(t,\beta^o,\hat\pi,S^*,G^o)-\tilde\Lambda_0(t,\beta^o,\pi^*,S^*,G^o) \qquad (4.44)\\
&=\int_0^t\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat\pi,S^*,G^o)-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)+\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n}\sum_{i=1}^n\Big(\left(\hat w_i-w_i^*\right)\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)-d\Lambda_0^o(u)\exp(\beta^oA_i)R_i(u,S^*,G^o)\right\}\\
&\quad-\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\quad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n}\sum_{j=1}^n\left(\hat w_j-w_j^*\right)\exp(\beta^oA_j)R_j(u,S^*,G^o)\Big).
\end{aligned}$$

$$
\begin{aligned}
&\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right) \qquad (4.45)\\
&=\int_0^t\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G})-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)+\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n}\sum_{i=1}^{n}\left(w_i^*K_i(u,A_i)-\sum_{a=0,1}K_i(u,a)+d\Lambda_0^o(u)\left\{w_i^*e^{\beta^oA_i}J_i(u,A_i)-\sum_{a=0,1}e^{\beta^oa}J_i(u,a)\right\}\right.\\
&\quad-\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\quad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\left.\times\frac{1}{n}\sum_{j=1}^{n}\left\{-w_j^*e^{\beta^oA_j}J_j(u,A_j)+\sum_{a=0,1}e^{\beta^oa}J_j(u,a)\right\}\right).
\end{aligned}
$$

where

$$K_i(u,A_i)=\hat{G}(u,A_i)d\hat{S}_i(u,A_i)-G^o(u,A_i)dS_i^*(u,A_i),$$

and

$$J_i(u,A_i)=\hat{G}(u,A_i)\hat{S}_i(u,A_i)-G^o(u,A_i)S_i^*(u,A_i).$$

The proof of Lemma 23 is reported in Section 4.8.4. Moreover, the technical Lemma 27-30 needed are reported in Section 4.8.5

Proof of Lemma 22. By algebra we have:

$$\begin{aligned}
U_{1,n}^{AIPW}\left(\beta^o;\hat{\pi},\hat{S},\hat{G}\right) &= U_{1,n}^{AIPW}\left(\beta^o;\pi^*,S^*,G^o\right) \\
&\quad + U_{1,n}^{AIPW}\left(\beta^o;\hat{\pi},\hat{S},\hat{G}\right) - U_{1,n}^{AIPW}\left(\beta^o;\pi^*,\hat{S},\hat{G}\right) \\
&\quad + U_{1,n}^{AIPW}\left(\beta^o;\pi^*,\hat{S},\hat{G}\right) - U_{1,n}^{AIPW}\left(\beta^o;\pi^*,S^*,G^o\right) \\
&= U_{1,n}^{AIPW}\left(\beta^o;\pi^*,S^*,G^o\right) + Q_1 + Q_2.
\end{aligned} \tag{4.46}$$

We now work on Q_1, Q_2 separately.

- Term Q_1:

By Cauchy-Schwartz and by the fact that $b_n c_n = o_p(n^{-1/2})$ we have:

$$Q_1 = U_{1,n}^{AIPW}\left(\beta^o;\hat{\pi},S^*,G^o\right) - U_{1,n}^{AIPW}\left(\beta^o;\pi^*,S^*,G^o\right) + o_p(n^{-1/2}).$$

Moreover, we have:

$$\begin{aligned}
Q_1 = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \Bigg[& \left(\frac{1}{\hat{\pi}_i} - \frac{1}{\pi_i^*}\right) A_i \left\{dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i)\right\} \\
& - \left\{\frac{1}{\hat{\pi}_i} d\tilde{\Lambda}_0\left(t;\beta^o,\hat{\pi},S^*,G^o\right) - \frac{1}{\pi_i^*} d\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\} A_i e^{\beta^o A_i} R_i(t,S^*,G^o) \\
& - d\left\{\tilde{\Lambda}_0\left(t;\beta^o,\hat{\pi},S^*,G^o\right) - \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\} e^{\beta^o} S_i^*(t,1) G^o(t,1) \Bigg].
\end{aligned}$$

By algebra we have:

$$\begin{aligned}
Q_1 &= \int_0^\tau \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_i^*}\right)A_i\left\{dN_i(t)+G^o(t,A_i)dS_i^*(t,A_i)\right.\\
&\qquad\left.-d\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)e^{\beta^o A_i}R_i(t,S^*,G^o)\right\}\\
&\quad-\int_0^\tau d\left\{\tilde{\Lambda}_0\left(t;\beta^o,\hat{\pi},S^*,G^o\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\}\mathcal{S}^{(1)}(t;\beta^o,\hat{\pi},S^*,G^o)\\
&= Q_{11}+Q_{12}+Q_{13},
\end{aligned}$$

where

$$\begin{aligned}
Q_{11} = \int_0^\tau \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_i^*}\right)A_i&\left[dN_i(t)+G^o(t,A_i)dS_i^*(t,A_i)-d\Lambda_0^o(t)e^{\beta^o A_i}R_i(t,S^*,G^o)\right.\\
&\left.-d\left\{\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)-\Lambda_0^o(t)\right\}e^{\beta^o A_i}R_i(t,S^*,G^o)\right],
\end{aligned}$$

$$Q_{12} = -\int_0^\tau d\left\{\tilde{\Lambda}_0\left(t;\beta^o,\hat{\pi},S^*,G^o\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\}\mathcal{S}^{(1)}(t;\beta^o,\pi^*,S^*,G^o),$$

and

$$\begin{aligned}
Q_{13} = -\int_0^\tau &d\left\{\tilde{\Lambda}_0\left(t;\beta^o,\hat{\pi},S^*,G^o\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\}\\
&\times\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_i^*}\right)A_i e^{\beta^o A_i}R_i(t,S^*,G^o).
\end{aligned}$$

By Lemma 23, we have:

$$
\begin{aligned}
Q_{11} = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{\hat{\pi}_i} - \frac{1}{\pi_i^*}\right) A_i \Bigg(& dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)e^{\beta^o A_i}R_i(t,S^*,G^o) \\
& - e^{\beta^o A_i}R_i(t,S^*,G^o)\frac{1}{n}\sum_{j=1}^n \Bigg[w_j^*\left\{dN_j(t) + G^o(t,A_j)dS_j^*(t,A_j)\right\} - \sum_{a=0,1} G^o(t,a)dS_j^*(t,a) \\
& - d\Lambda_0^o(t)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\Bigg] \left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\Bigg),
\end{aligned}
$$

$$
\begin{aligned}
Q_{12} = & -\int_0^\tau \mathcal{S}^{(1)}(t;\beta^o,\pi^*,S^*,G^o) \\
& \times \left\{\mathcal{S}^{(0)}(t;\beta^o,\hat{\pi},S^*,G^o) - \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) + \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
& \times \frac{1}{n}\sum_{i=1}^n \left((\hat{w}_i - w_i^*)\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)\exp(\beta^o A_i)R_i(t,S^*,G^o)\right\}\right. \\
& - \Bigg[w_i^*\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i)\right\} - \sum_{a=0,1} G^o(t,a)dS_i^*(t,a) \\
& \quad - d\Lambda_0^o(t)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\Bigg] \left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
& \quad \left. \times \frac{1}{n}\sum_{j=1}^n (\hat{w}_j - w_j^*)\exp(\beta^o A_j)R_j(t,S^*,G^o)\right),
\end{aligned}
$$

and

$$
\begin{aligned}
Q_{13} = &-\int_0^\tau \frac{1}{n}\sum_{j=1}^n \left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j^*}\right) A_j e^{\beta^o A_j} R_j(t, S^*, G^o) \\
&\times \left\{ \mathcal{S}^{(0)}(t;\beta^o,\hat{\pi},S^*,G^o) - \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) + \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) \right\}^{-1} \\
&\times \frac{1}{n}\sum_{i=1}^n \left((\hat{w}_i - w_i^*) \left\{ dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i) - d\Lambda_0^o(t) \exp(\beta^o A_i) R_i(t,S^*,G^o) \right\} \right. \\
&- \left[w_i^* \left\{ dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i) \right\} - \sum_{a=0,1} G^o(t,a) dS_i^*(t,a) \right. \\
&\quad \left. -d\Lambda_0^o(t) \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) \right] \left\{ \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) \right\}^{-1} \\
&\quad \left. \times \frac{1}{n}\sum_{j=1}^n (\hat{w}_j - w_j^*) \exp(\beta^o A_j) R_j(t,S^*,G^o) \right).
\end{aligned}
$$

By Assumptions 7 and 8 we have:

$$
\begin{aligned}
\sup_{t\in[0,\tau]} \left| \frac{1}{n}\sum_{j=1}^n (\hat{w}_j - w_j^*) \exp(\beta^o A_j) R_j(t,S^*,G^o) \right| &\leq 2\exp(\beta^o) \left| \frac{1}{n}\sum_{j=1}^n (\hat{w}_j - w_j^*) \right| \\
&= o_p(1),
\end{aligned}
$$

and similarly

$$
\sup_{t\in[0,\tau]} \left| \frac{1}{n}\sum_{j=1}^n (\hat{w}_j - w_j^*) A \exp(\beta^o A_j) R_j(t,S^*,G^o) \right| = o_p(1).
$$

Therefore by Lemma 28 and Assumptions 7 and 9, we have:

$$
\begin{aligned}
Q_{11} = &\int_0^\tau \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{\hat{\pi}_i} - \frac{1}{\pi_i^*}\right) A_i \left(dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i) - d\Lambda_0^o(t) e^{\beta^o A_i} R_i(t,S^*,G^o) \right) \\
&+ o_p(n^{-1/2}),
\end{aligned}
$$

$$
\begin{aligned}
Q_{12} = &-\int_0^\tau s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
&\times \frac{1}{n}\sum_{i=1}^n (\hat{w}_i - w_i^*)\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)\exp(\beta^o A_i)R_i(t,S^*,G^o)\right\} \\
&+ o_p(n^{-1/2}),
\end{aligned}
$$

and

$$
\begin{aligned}
Q_{13} = &-\int_0^\tau \frac{1}{n}\sum_{j=1}^n \left(\frac{1}{\hat{\pi}_j} - \frac{1}{\pi_j^*}\right) A_j e^{\beta^o A_j} R_j(t,S^*,G^o)\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
&\times \frac{1}{n}\sum_{i=1}^n (\hat{w}_i - w_i^*)\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)\exp(\beta^o A_i)R_i(t,S^*,G^o)\right\} \\
&+ o_p(n^{-1/2}).
\end{aligned}
$$

By the fact that $\sup_{Z\in\mathcal{Z}}|\hat{\pi}(Z) - \pi^*(Z)| \to 0$, it is now easy to see that $Q_1 = o_p(1)$.

We now divide the proof in two parts according to the three scenarios a,b,c of the Lemma.

- Case a) and c) of the Lemma: $S^* = S^o$.

Since $S^* = S^o$, by Lemma 27, we have $Q_{11} = o_p(n^{-1/2})$, $Q_{12} = o_p(n^{-1/2})$ and $Q_{13} = o_p(n^{-1/2})$. Therefore $Q_1 = o_p(n^{-1/2})$.

- Case b) of the Lemma: $S^* \neq S^o$, $\pi^* = \pi^o$ with $b_n = n^{-1/2}$.

Plugging into Q_{11}, Q_{12}, Q_{13} the influence function of $\hat{\pi} - \pi^o$, we have:

$$\begin{aligned}
&Q_{11} \\
&= -\int_0^\tau \frac{1}{n^2} \sum_{i,j=1}^n \phi_j(Z_i) \{\pi^*(Z_i)\}^{-2} A_i \left(dN_i(t) + G^o(t, A_i) dS_i^*(t, A_i) - d\Lambda_0^o(t) e^{\beta^o A_i} R_i(t, S^*, G^o) \right) \\
&\quad + o_p(n^{-1/2}),
\end{aligned}$$

$$\begin{aligned}
Q_{12} = &-\int_0^\tau s^{(1)}(t; \beta^o, \pi^*, S^*, G^o) \left\{ s^{(0)}(t; \beta^o, \hat{\pi}, S^*, G^o) \right\}^{-1} \\
&\times \frac{1}{n^2} \sum_{i,j=1}^n \phi_j(Z_i) \left[\frac{1 - A_i}{\left\{1 - \pi_i^*\right\}^2} - \frac{A_i}{\left\{\pi_i^*\right\}^2} \right] \\
&\times \{ dN_i(t) + G^o(t, A_i) dS_i^*(t, A_i) - d\Lambda_0^o(t) \exp(\beta^o A_i) R_i(t, S^*, G^o) \} + o_p(n^{-1/2}),
\end{aligned}$$

$$\begin{aligned}
Q_{13} = &\int_0^\tau \frac{1}{n^2} \sum_{i,j=1}^n \phi_i(Z_j) \left\{\pi^*(Z_j)\right\}^{-2} A_j e^{\beta^o A_j} R_j(t, S^*, G^o) \left\{ s^{(0)}(t; \beta^o, \pi^*, S^*, G^o) \right\}^{-1} \\
&\times \frac{1}{n^2} \sum_{l,m=1}^n \phi_l(Z_m) \left[\frac{1 - A_l}{\left\{1 - \pi_l^*\right\}^2} - \frac{A_l}{\left\{\pi_l^*\right\}^2} \right] \\
&\times \{ dN_l(t) + G^o(t, A_l) dS_l^*(t, A_l) - d\Lambda_0^o(t) \exp(\beta^o A_l) R_l(t, S^*, G^o) \} \\
&+ o_p(n^{-1/2}).
\end{aligned}$$

We can therefore conclude the following:

$$
\begin{aligned}
Q_1 = &-\int_0^\tau \frac{1}{n^2}\sum_{i,j=1}^{n} \phi_j(Z_i)\left\{\pi_i^*\right\}^{-2} A_i \left(dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)e^{\beta^o A_i}R_i(t,S^*,G^o)\right)\\
&-\int_0^\tau s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)\left\{s^{(0)}(t;\beta^o,\hat{\pi},S^*,G^o)\right\}^{-1}\frac{1}{n^2}\sum_{i,j=1}^{n}\phi_j(Z_i)\left[\frac{1-A_i}{\left\{1-\pi_i^*\right\}^2} - \frac{A_i}{\left\{\pi_i^*\right\}^2}\right]\\
&\quad\times\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i) - d\Lambda_0^o(t)\exp(\beta^o A_i)R_i(t,S^*,G^o)\right\}\\
&+\int_0^\tau \frac{1}{n^2}\sum_{i,j=1}^{n}\phi_i(Z_j)\left\{\pi_j^*\right\}^{-2}A_j e^{\beta^o A_j}R_j(t,S^*,G^o)\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n^2}\sum_{l,m=1}^{n}\phi_l(Z_m)\left[\frac{1-A_l}{\left\{1-\pi_l^*\right\}^2} - \frac{A_l}{\left\{\pi_l^*\right\}^2}\right]\\
&\quad\times\left\{dN_l(t) + G^o(t,A_l)dS_l^*(t,A_l) - d\Lambda_0^o(t)\exp(\beta^o A_l)R_l(t,S^*,G^o)\right\}\\
&+o_p(n^{-1/2}).
\end{aligned}
$$

- Term Q_2:

For ease of exposition we define:

$$K_i(t,A_i) = \hat{G}(t,A_i)d\hat{S}_i(t,A_i) - G^o(t,A_i)dS_i^*(t,A_i),$$

and

$$J_i(t,A_i) = \hat{G}(t,A_i)\hat{S}_i(t,A_i) - G^o(t,A_i)S_i^*(t,A_i).$$

By definition we have:

$$
\begin{aligned}
&Q_2 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.47)\\
&= \int_0^\tau \frac{1}{n}\sum_{i=1}^n \left(\frac{A_i}{\pi_i^*} \left[-d\left\{ \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\} e^{\beta^o A_i} Y_i(t) + K_i(t,A_i) \right.\right.\\
&\left. + e^{\beta^o A_i}\left\{ d\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right)\hat{S}_i(t,A_i)\hat{G}(t,A_i) - d\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right) S_i^*(t,A_i)G^o(t,A_i)\right\}\right]\\
&\left. - K_i(t,1) - e^{\beta^o}\left\{ d\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right)\hat{S}_i(t,1)\hat{G}(t,1) - d\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right) S_i^*(t,1)G^o(t,1)\right\}\right).
\end{aligned}
$$

Algebra gives us:

$$
\begin{aligned}
Q_2 = \int_0^\tau \frac{1}{n}\sum_{i=1}^n & \left[-d\left\{ \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\} \mathcal{S}^{(1)}(t;\beta^o,\pi^*,S^*,G^o) \right.\\
& + d\Lambda_0^o(t)\left\{ \frac{A_i}{\pi_i^*} J_i(t,A_i) e^{\beta^o A_i} - J_i(t,a)e^{\beta^o}\right\}\\
& + d\left\{ \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \Lambda_0^o(t)\right\} \left\{ J_i(t,A_i)\frac{A_i}{\pi_i^*} e^{\beta^o A_i} - J_i(t,a)e^{\beta^o}\right\}\\
& \left. + \frac{A_i}{\pi_i^*} K_i(u,A_i) - K_i(u,1)\right]\\
= {} & Q_{21} + Q_{22} + Q_{23} + Q_{24},
\end{aligned}
$$

where

$$
Q_{21} = -\int_0^\tau \frac{1}{n}\sum_{i=1}^n d\left\{ \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\right\} \mathcal{S}^{(1)}(t;\beta^o,\pi^*,S^*,G^o),
$$

$$
Q_{22} = \int_0^\tau d\Lambda_0^o(t)\frac{1}{n}\sum_{i=1}^n \left[J_i(t,1)\left\{ \frac{A_i}{\pi_i^*} e^{\beta^o A_i} - e^{\beta^o}\right\}\right],
$$

$$
\begin{aligned}
Q_{23} = \int_0^\tau d\left\{\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right) + \tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right) - \Lambda_0^o(t)\right\} \\
\times \frac{1}{n}\sum_{i=1}^n \left[J_i(t,1)\left\{\frac{A_i}{\pi_i^*}e^{\beta^o A_i} - e^{\beta^o}\right\}\right],
\end{aligned}
$$

and

$$
Q_{24} = \int_0^\tau \frac{1}{n}\sum_{i=1}^n K_i(t,1)\left\{\frac{A_i}{\pi_i^*} - 1\right\}.
$$

By Lemma 23 we have:

$$
\begin{aligned}
Q_{21} = &-\int_0^\tau \mathcal{S}^{(1)}(t;\beta^o,\pi^*,S^*,G^o) \\
&\times \left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) + \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
&\times \frac{1}{n}\sum_{i=1}^n \left(w_i^* K_i(t,A_i) - \sum_{a=0,1} K_i(t,a) + d\Lambda_0^o(t)\left\{w_i^* e^{\beta^o A_i} J_i(u,A_i) - \sum_{a=0,1} e^{\beta^o a} J_i(t,a)\right\}\right. \\
&\quad - \left[w_i^*\left\{dN_i(t) + G^o(t,A_i)dS_i^*(t,A_i)\right\} - \sum_{a=0,1} G^o(t,a)dS_i^*(t,a)\right. \\
&\quad \left. - d\Lambda_0^o(t)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
&\quad \left. \times \frac{1}{n}\sum_{j=1}^n \left\{-w_j^* e^{\beta^o A_j} J_j(t,A_j) + \sum_{a=0,1} e^{\beta^o a} J_j(t,a)\right\}\right),
\end{aligned}
$$

$$\begin{aligned}
Q_{23} = \int_0^\tau \frac{1}{n}\sum_{j=1}^n & \left[J_j(t,1)\left\{\frac{A_j}{\pi_j^*}e^{\beta^o A_j} - e^{\beta^o}\right\}\right] \\
& \times\left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o) + \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
& \times \frac{1}{n}\sum_{i=1}^n \Bigg(w_i^* K_i(t,A_i) - \sum_{a=0,1} K_i(t,a) + d\Lambda_0^o(u)\left\{ w_i^* e^{\beta^o A_i} J_i(t,A_i) - \sum_{a=0,1} e^{\beta^o a} J_i(t,a)\right\} \\
& \quad - \Bigg[w_i^* \left\{ dN_i(t) + G^o(t,A_i) dS_i^*(t,A_i)\right\} - \sum_{a=0,1} G^o(t,a) dS_i^*(t,a) \\
& \quad - d\Lambda_0^o(u)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\Bigg] \left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
& \quad \times \frac{1}{n}\sum_{l=1}^n \left\{ -w_l^* e^{\beta^o A_l} J_l(t,A_l) + \sum_{a=0,1} e^{\beta^o a} J_l(t,a)\right\}\Bigg) \\
& + \int_0^\tau \frac{1}{n}\sum_{i=1}^n \left[J_i(t,1)\left\{\frac{A_i}{\pi_i^*}e^{\beta^o A_i} - e^{\beta^o}\right\}\right] \\
& \quad \times \Bigg\{ \frac{1}{n}\sum_{j=1}^n \left[w_j^* \left\{ dN_j(t) + G^o(t,A_j) dS_j^*(t,A_j)\right\} - \sum_{a=0,1} G^o(t,a) dS_j^*(t,a)\right] \\
& \quad - d\Lambda_0^o(u)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\Bigg\} \left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}.
\end{aligned}$$

By Assumption 7 we have, for $a = 0,1$:

$$\sup_{t\in[0,\tau], i=1,\ldots,n} |J_i(t,a)| = o_p(1). \tag{4.48}$$

Therefore, by Lemma 28 and by Assumption 9, we have:

$$\begin{aligned}
Q_{21} = -\int_0^\tau & s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
& \times \frac{1}{n}\sum_{i=1}^n \left(w_i^* K_i(t,A_i) - \sum_{a=0,1} K_i(t,a) + d\Lambda_0^o(t)\left\{ w_i^* e^{\beta^o A_i} J_i(u,A_i) - \sum_{a=0,1} e^{\beta^o a} J_i(t,a)\right\}\right) \\
& + o_p(n^{-1/2}),
\end{aligned}$$

$$Q_{23} = \int_0^\tau \frac{1}{n}\sum_{j=1}^n \left[J_j(t,1)\left\{ \frac{A_j}{\pi_j^*} e^{\beta^o A_j} - e^{\beta^o} \right\}\right] \left\{ s^{(0)}(t;\beta^o,\pi^*,S^*,G^o) \right\}^{-1}$$
$$\times \frac{1}{n}\sum_{i=1}^n \left(w_i^* K_i(t,A_i) - \sum_{a=0,1} K_i(t,a) + d\Lambda_0^o(t) \left\{ w_i^* e^{\beta^o A_i} J_i(t,A_i) - \sum_{a=0,1} e^{\beta^o a} J_i(t,a) \right\}\right)$$
$$+ o_p(n^{-1/2}).$$

By Assumption 7, it is easy to see that $Q_2 = o_p(1)$.

We now divide the proof into two parts according to the three scenarios a,b,c of the Lemma.

- Case a) and b) of the Lemma: $\pi^* = \pi^o$.

By Assumption 7 we have, for $a = 0,1$:

$$\sup_{i=1,\ldots,n} \int_0^\tau |K_i(t,a)| = o_p(1). \tag{4.49}$$

Since $\pi^* = \pi^0$, by Lemma 29 and by (4.49) we can conclude that $Q_{22} = o_p(n^{-1/2})$, $Q_{24} = o_p(n^{-1/2})$. Moreover, by Lemma 30 and by (4.48) and (4.49), we can conclude that $Q_{21} = o_p(n^{-1/2})$ and $Q_{23} = o_p(n^{-1/2})$.

- Case c) of the Lemma: $\pi^* \neq \pi^o$, $S^* = S^o$ with $c_n = o(n^{-1/2})$.

We notice that, we have:

$$K_i(t,A_i) = \frac{1}{n}\sum_{j=1}^n \left\{ G^o(t,A_i) d\psi_j(t,A_i,Z_i) + dS_i^o(t,A_i)\rho_j(t,A_i) \right\} + o_p(n^{-1/2}), \tag{4.50}$$

and

$$J_i(t,A_i) = \frac{1}{n}\sum_{j=1}^n \left\{ G^o(t,A_i) \psi_j(t,A_i,Z_i) + S_i^o(t,A_i)\rho_j(t,A_i) \right\} + o_p(n^{-1/2}). \tag{4.51}$$

Plugging (4.50) and (4.51) into $Q_{21}, Q_{22}, Q_{23}, Q_{24}$ we have:

$$\begin{aligned}
Q_{21} = &-\int_0^\tau s^{(1)}(t;\beta^o,\pi^*,S^*,G^o)\left\{s^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
&\times \frac{1}{n^2}\sum_{i,j=1}^n \Big(w_i^*\left\{G^o(t,A_i)d\psi_j(t,A_i,Z_i)+dS_i^o(t,A_i)\rho_j(t,A_i)\right\} \\
&- \sum_{a=0,1}\left\{G^o(t,a)d\psi_j(t,a,Z_i)+dS_i^o(t,a)\rho_j(t,a)\right\} \\
&+ d\Lambda_0^o(t)\Big\{w_i^* e^{\beta^o A_i}\left\{G^o(t,A_i)\psi_j(t,A_i,Z_i)+S_i^o(t,A_i)\rho_j(t,A_i)\right\} \\
&- \sum_{a=0,1} e^{\beta^o a}\left\{G^o(t,a)\psi_j(t,a,Z_i)+S_i^o(t,a)\rho_j(t,a)\right\}\Big) \\
&+ o_p(n^{-1/2}),
\end{aligned}$$

$$Q_{22} = \int_0^\tau d\Lambda_0^o(t)\frac{1}{n^2}\sum_{i,j=1}^n \left\{G^o(t,1)\psi_j(t,1,Z_i)+S_i^o(t,1)\rho_j(t,1)\right\}\left\{\frac{A_i}{\pi_i^*}e^{\beta^o A_i}-e^{\beta^o}\right\},$$

$$\begin{aligned}
Q_{23} = &\int_0^\tau \frac{1}{n}\sum_{l=1}^n\left[J_l(t,1)\left\{\frac{A_l}{\pi_l^*}e^{\beta^o A_l}-e^{\beta^o}\right\}\right]\left\{s^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
&\times \frac{1}{n^2}\sum_{i,j=1}^n \Big(w_i^*\left\{G^o(t,A_i)d\psi_j(t,A_i,Z_i)+dS_i^o(t,A_i)\rho_j(t,A_i)\right\} \\
&- \sum_{a=0,1}\left\{G^o(t,a)d\psi_j(t,a,Z_i)+dS_i^o(t,a)\rho_j(t,a)\right\} \\
&+ d\Lambda_0^o(t)\Big[w_i^* e^{\beta^o A_i}\left\{G^o(t,A_i)\psi_j(t,A_i,Z_i)+S_i^o(t,A_i)\rho_j(t,A_i)\right\} \\
&- \sum_{a=0,1} e^{\beta^o a}\left\{G^o(t,a)\psi_j(t,a,Z_i)+S_i^o(t,a)\rho_j(t,a)\right\}\Big]\Big) \\
&+ o_p(n^{-1/2}),
\end{aligned}$$

$$Q_{24} = \int_0^\tau \frac{1}{n^2} \sum_{i,j=1}^n \left\{ G^o(t,A_i) d\psi_j(t,1,Z_i) + dS_i^o(t,A_i)\rho_j(t,1) \right\} \left\{ \frac{A_i}{\pi_i^*} - 1 \right\} + o_p(n^{-1/2}).$$

Therefore, we can conclude that:

$$\begin{aligned}
Q_2 = & -\int_0^\tau \left\{ s^{(0)}(t;\beta^o,\pi^*,S^*,G^o) \right\}^{-1} \left[s^{(1)}(t;\beta^o,\pi^*,S^*,G^o) + \frac{1}{n}\sum_{l=1}^n J_l(t,1) \left\{ \frac{A_l}{\pi_l^*} e^{\beta^o A_l} - e^{\beta^o} \right\} \right] \\
& \quad \times \frac{1}{n^2} \sum_{i,j=1}^n \Big(w_i^* \left\{ G^o(t,A_i) d\psi_j(t,A_i,Z_i) + dS_i^o(t,A_i)\rho_j(t,A_i) \right\} \\
& \qquad - \sum_{a=0,1} \left\{ G^o(t,a) d\psi_j(t,a,Z_i) + dS_i^o(t,a)\rho_j(t,a) \right\} \\
& \qquad + d\Lambda_0^o(u) \Big\{ w_i^* e^{\beta^o A_i} \left\{ G^o(t,A_i)\psi_j(t,A_i,Z_i) + S_i^o(t,A_i)\rho_j(t,A_i) \right\} \\
& \qquad - \sum_{a=0,1} e^{\beta^o a} \left\{ G^o(t,a)\psi_j(t,a,Z_i) + S_i^o(t,a)\rho_j(t,a) \right\} \Big) \\
& + \int_0^\tau d\Lambda_0^o(t) \frac{1}{n^2} \sum_{i,j=1}^n \left\{ G^o(t,1)\psi_j(t,1,Z_i) + S_i^o(t,1)\rho_j(t,1) \right\} \left\{ \frac{A_i}{\pi_i^*} e^{\beta^o A_i} - e^{\beta^o} \right\} \\
& + \int_0^\tau \frac{1}{n^2} \sum_{i,j=1}^n \left\{ G^o(t,A_i) d\psi_j(t,1,Z_i) + dS_i^o(t,A_i)\rho_j(t,1) \right\} \left\{ \frac{A_i}{\pi_i^*} - 1 \right\}.
\end{aligned}$$

Therefore, putting together all the results, the lemma is proved. □

4.8.4 Proofs of additional Lemmas

Proof of Lemma 23. We remind the reader that, for each $t \in [0,\tau]$, $\tilde{\Lambda}_0(t;\beta^o,\pi,S,G)$ is the root of $U_{2,n}^{AIPW}(t,\Lambda_0;\beta^o,\pi,S,G) = 0$. The score $U_{2,n}^{AIPW}(t,\Lambda_0;\beta^o,\pi,S,G)$ can be written as:

$$\begin{aligned} U_{2,n}^{AIPW}(t,\Lambda_0;\beta^o,\pi,S,G) &= \int_0^t V_1(u;\pi,S,G) - \int_0^t d\Lambda_0(u)\mathcal{S}^{(0)}(u;\beta^o,\pi,S,G) \\ &= \int_0^t V_1(u;\pi,S,G) - \int_0^t d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi,S,G) \\ &\quad - \int_0^t d\left\{\Lambda_0(u) - \Lambda_0^o(u)\right\}\mathcal{S}^{(0)}(u;\beta^o,\pi,S,G), \end{aligned} \tag{4.52}$$

where

$$V_1(u;\pi,S,G) = \frac{1}{n}\sum_{i=1}^{n}\left[w_i\left\{dN_i(u) + G(u,A_i)dS_i(u,A_i)\right\} - \sum_{a=0,1} G(u,a)dS_i(u,a)\right].$$

- Proof of (4.43):

By (4.52), we have, for each $t \in [0,\tau]$:

$$\begin{aligned} &\tilde{\Lambda}_0(t;\beta^o,\pi^*,S^*,G^o) - \Lambda_0^o(t) \\ &= \int_0^t \left\{V_1(u;\pi^*,S^*,G^o) - d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\ &= \int_0^t \frac{1}{n}\sum_{i=1}^{n}\left[w_i^*\left\{dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i)\right\} - \sum_{a=0,1} G^o(u,a)dS_i^*(u,a)\right. \\ &\qquad \left. - d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}. \end{aligned}$$

- Proof of (4.44):

By (4.52) we have:

$$\begin{aligned}
0 &= U_{2,n}^{AIPW}\left\{\tilde{\Lambda}_0\left(t;\beta^o,\hat{\pi},S^*,G^o\right);\beta^o,\hat{\pi},S^*,G^o\right\} - U_{2,n}^{AIPW}\left\{\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right);\beta^o,\pi^*,S^*,G^o\right\} \\
&= \int_0^t \{V_1(u;\hat{\pi},S^*,G^o) - V_1(u;\pi^*,S^*,G^o)\} \\
&\quad - \int_0^t d\tilde{\Lambda}_0\left(u;\beta^o,\hat{\pi},S^*,G^o\right)\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^o) \\
&\quad + \int_0^t d\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \\
&= \int_0^t \frac{1}{n}\sum_{i=1}^n \left(\hat{w}_i - w_i^*\right)\{dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i)\} \\
&\quad - \int_0^t d\tilde{\Lambda}_0\left(u;\beta^o,\hat{\pi},S^*,G^o\right)\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^o) \\
&\quad + \int_0^t d\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o).
\end{aligned}$$

We have:

$$\begin{aligned}
0 = {} & \frac{1}{n}\sum_{i=1}^n \int_0^t \left(\hat{w}_i - w_i^*\right)\{dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i)\} \\
& - \int_0^t d\left\{\tilde{\Lambda}_0(u,\beta^o,\hat{\pi},S^*,G^o) - \tilde{\Lambda}_0(u,\beta^o,\pi^*,S^*,G^o)\right\}\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0) \\
& - \int_0^t d\tilde{\Lambda}_0(u,\beta^o,\hat{\pi},S^*,G^o)\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\}.
\end{aligned}$$

Moreover:

$$
\begin{aligned}
0 = \frac{1}{n}\sum_{i=1}^{n}\int_0^t & (\hat{w}_i - w_i^*)\left\{dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i)\right\} \\
& - \int_0^t d\left\{\tilde{\Lambda}_0(u,\beta^o,\hat{\pi},S^*,G^o) - \tilde{\Lambda}_0(u,\beta^o,\pi^*,S^*,G^o)\right\}\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0) \\
& - \int_0^t d\Lambda_0^o(u)\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\} \\
& - \int_0^t d\left\{\tilde{\Lambda}_0(u,\beta^o,\hat{\pi},S^*,G^o) - \tilde{\Lambda}_0(u,\beta^o,\pi^*,S^*,G^o)\right\} \\
& \qquad \times\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\} \\
& - \int_0^t d\left\{\tilde{\Lambda}_0(u,\beta^o,\pi^*,S^*,G^o) - \Lambda_0^o(u)\right\}\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\}.
\end{aligned}
$$

By (4.43), we therefore have:

$$
\begin{aligned}
0 = \frac{1}{n}\sum_{i=1}^{n}\int_0^t & (\hat{w}_i - w_i^*)\left\{dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i)\right\} \\
& - \int_0^t d\left\{\tilde{\Lambda}_0(u,\beta^o,\hat{\pi},S^*,G^o) - \tilde{\Lambda}_0(u,\beta^o,\pi^*,S^*,G^o)\right\}\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0) \\
& - \int_0^t d\Lambda_0^o(u)\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\} \\
& - \int_0^t d\left\{\tilde{\Lambda}_0(u,\beta^o,\hat{\pi},S^*,G^o) - \tilde{\Lambda}_0(u,\beta^o,\pi^*,S^*,G^o)\right\} \\
& \qquad \times\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\} \\
& - \int_0^t \frac{1}{n}\sum_{i=1}^{n}\left[w_i^*\left\{dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i)\right\} - \sum_{a=0,1} G^o(u,a)dS_i^*(u,a)\right. \\
& \quad \left. - d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1} \\
& \qquad \times\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\}.
\end{aligned}
$$

Therefore, solving for $\tilde{\Lambda}_0(t,\beta^o,\hat{\pi},S^*,G^o)-\tilde{\Lambda}_0(t,\beta^o,\pi^*,S^*,G^o)$, we have:

$$
\begin{aligned}
&\tilde{\Lambda}_0(t,\beta^o,\hat{\pi},S^*,G^o)-\tilde{\Lambda}_0(t,\beta^o,\pi^*,S^*,G^o)\\
&\quad=\int_0^t\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^o)-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)+\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\qquad\times\frac{1}{n}\sum_{i=1}^{n}\left((\hat{w}_i-w_i^*)\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}\right.\\
&\qquad-d\Lambda_0^o(t)\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0)-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\}\\
&\qquad-\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\qquad\quad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\qquad\quad\left.\times\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^0)-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^0)\right\}\right).
\end{aligned}
$$

By definition of $\mathcal{S}^{(0)}$, we then have:

$$
\begin{aligned}
&\tilde{\Lambda}_0(t,\beta^o,\hat{\pi},S^*,G^o)-\tilde{\Lambda}_0(t,\beta^o,\pi^*,S^*,G^o)\\
&\quad=\int_0^t\left\{\mathcal{S}^{(0)}(u;\beta^o,\hat{\pi},S^*,G^o)-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)+\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n}\sum_{i=1}^{n}\left((\hat{w}_i-w_i^*)\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)-d\Lambda_0^o(u)\exp(\beta^oA_i)R_i(u,S^*,G^o)\right\}\right.\\
&\quad-\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\qquad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\qquad\left.\times\frac{1}{n}\sum_{j=1}^{n}(\hat{w}_j-w_j^*)\exp(\beta^oA_j)R_j(u,S^*,G^o)\right).
\end{aligned}
$$

- Proof of (4.45):

By (4.52), we have:

$$
\begin{aligned}
0 &= U_2\left(\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right);\beta^o,\pi^*,\hat{S},\hat{G}\right) - U_2\left(\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right);\beta^o,\pi^*,S^*,G^o\right) \\
&= \int_0^t \left\{V_1(u;\pi^*,\hat{S},\hat{G}) - V_1(u;\pi^*,S^*,G^o)\right\} - \int_0^t d\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,\hat{S},\hat{G}\right)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) \\
&\quad + \int_0^t d\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o).
\end{aligned}
$$

Tedious algebra gives us:

$$
\begin{aligned}
0 &= U_2\left(\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right);\beta^o,\pi^*,\hat{S},\hat{G}\right) - U_2\left(\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right);\beta^o,\pi^*,S^*,G^o\right) \\
&= \int_0^t \frac{1}{n}\sum_{i=1}^n \left\{w_i^* K_i(u,A_i) - \sum_{a=0,1} K_i(u,a)\right\} \\
&\quad - \int_0^t d\Lambda_0^o(u)\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\} \\
&\quad - \int_0^t d\left\{\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right)\right\}\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \\
&\quad - \int_0^t d\left\{\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right)\right\} \\
&\qquad\quad \times \left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\} \\
&\quad - \int_0^t d\left\{\tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right) - \Lambda_0^o(u)\right\}\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\},
\end{aligned}
$$

where

$$
K_i(u,A_i) = \hat{G}(u,A_i)d\hat{S}_i(u,A_i) - G^o(u,A_i)dS_i^*(u,A_i).
$$

Therefore, by (4.43), we get:

$$\begin{aligned}
0 &= U_2\left(\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right);\beta^o,\pi^*,\hat{S},\hat{G}\right) - U_2\left(\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right);\beta^o,\pi^*,S^*,G^o\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{ w_i^* K_i(u,A_i) - \sum_{a=0,1} K_i(u,a) \right\} \\
&\quad - \int_0^t d\Lambda_0^o(u)\left\{ \mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \right\} \\
&\quad - \int_0^t d\left\{ \tilde{\Lambda}_0\left(u;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right) \right\} \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \\
&\quad - \int_0^t d\left\{ \tilde{\Lambda}_0\left(u;\beta^o,\pi^*,\hat{S},\hat{G}\right) - \tilde{\Lambda}_0\left(u;\beta^o,\pi^*,S^*,G^o\right) \right\} \\
&\qquad\quad \times \left\{ \mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \right\} \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{ w_i^*\left\{ dN_i(u) + G^o(u,A_i)dS_i^*(u,A_i) \right\} - \sum_{a=0,1} G^o(u,a)dS_i^*(u,a) \right. \\
&\qquad \left. - d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \right\} \left\{ \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \right\}^{-1} \\
&\qquad \times \left\{ \mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G}) - \mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o) \right\}.
\end{aligned}$$

Therefore, solving for $\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)$, we get:

$$
\begin{aligned}
&\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\\
&=\int_0^t\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G})-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)+\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n}\sum_{i=1}^{n}\left(w_i^*K_i(u,A_i)-\sum_{a=0,1}K_i(u,a)\right.\\
&\qquad-d\Lambda_0^o(u)\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G})-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}\\
&\qquad-\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\qquad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\qquad\left.\times\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G})-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}\right).
\end{aligned}
$$

We define:

$$J_i(u,A_i)=\hat{G}(u,A_i)\hat{S}_i(u,A_i)-G^o(u,A_i)S_i^*(u,A_i).$$

By definition of $\mathcal{S}^{(0)}$, we have:

$$\begin{aligned}
&\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,\hat{S},\hat{G}\right)-\tilde{\Lambda}_0\left(t;\beta^o,\pi^*,S^*,G^o\right)\\
&=\int_0^t\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,\hat{S},\hat{G})-\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)+\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\times\frac{1}{n}\sum_{i=1}^n\left(w_i^*K_i(u,A_i)-\sum_{a=0,1}K_i(u,a)+d\Lambda_0^o(u)\left\{w_i^*e^{\beta^oA_i}J_i(u,A_i)-\sum_{a=0,1}e^{\beta^oa}J_i(u,a)\right\}\right.\\
&\quad-\left[w_i^*\left\{dN_i(u)+G^o(u,A_i)dS_i^*(u,A_i)\right\}-\sum_{a=0,1}G^o(u,a)dS_i^*(u,a)\right.\\
&\quad\left.-d\Lambda_0^o(u)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right]\left\{\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}^{-1}\\
&\quad\left.\times\frac{1}{n}\sum_{j=1}^n\left\{-w_j^*e^{\beta^oA_j}J_j(u,A_j)+\sum_{a=0,1}e^{\beta^oa}J_j(u,a)\right\}\right).
\end{aligned}$$

□

4.8.5 Useful Lemmas and Results

Lemma 24 (Claim of Van Der Laan et al. (2003) (pag.87))**.** *If* $P(A=1|X(1),X(0),\delta(1),\delta(0),Z)=P(A=1|Z)$ *holds, the tangent space to the propensity score* $P(A=1\mid Z)$ *has the following form:* $\mathcal{T}=\{\phi(A,Z)\ :\ \textit{for all}\ \phi(A,Z)\ \textit{s.t}\ E\{\phi(A,Z)\mid Z\}=0\}$.

Lemma 25 (Lemma 1.4 of Van Der Laan et al. (2003))**.** *Suppose that* $Z=\psi(X)$ *where* X *is a random variable and* ψ *is a given function. Then, for any function q:* $\prod[q(X)\mid\{\phi(Z)\ :\ \textit{any}\ \phi\}]=E\{q(X)\mid Z\}$.

Lemma 26 (Lemma 1.5 of Van Der Laan et al. (2003))**.** *Let* (A,Z) *be a joint random variable. Let* $\mathcal{T}=\{\phi(A,Z)\ :\ \textit{for all}\ \phi(A,Z)\ \textit{s.t}\ E\{\phi(A,Z)\mid Z\}=0\}$. *Then, for any function q:* $\prod[q(A,Z)\mid\mathcal{T}]=q(A,Z)-E\{q(A,Z)\mid Z\}$.

Lemma 27. *For any $H(t,A,Z)$ such that $\sup_{[0,\tau],Z\in\mathcal{Z}}|H(t,a,Z)| = o_p(1)$ for $a = 0,1$, under Assumptions 10-6 we have:*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau H(t,A_i,Z_i)\left\{dN_i(t)+G^o(t,A_i)dS_i^o(t,A_i)-d\Lambda_0^o(t)e^{\beta^o A_i}R_i(t,S^o,G^o)\right\} = o_p(n^{-1/2}).$$

Proof of Lemma 27. Let's notice that, by (4.11) in the main document:

$$\begin{aligned} Q(t,A,Z) &:= dN(t)+G^o(t|A,Z)dS^o(t,A,Z)-d\Lambda_0^o(t)e^{\beta^o A}R(t,S^o,G^o) \\ &= dM(t)-d\mathrm{E}\{M(t)|A,Z\}. \end{aligned}$$

Therefore:

$$\begin{aligned} &\mathrm{E}\left\{\int_0^\tau H(t,A,Z)Q(t,A,Z)\right\} \\ &= \mathrm{E}\left[\int_0^\tau \mathrm{E}\{H(t,A,Z)dM(t)|A,Z\} - \int_0^\tau \mathrm{E}\{H(t,A,Z)dM(t)|A,Z\}\right] = 0. \end{aligned}$$

Moreover, for each i, there exists $C<\infty$ such that:

$$\left|\int_0^\tau H(t,A_i,Z_i)Q(t,A_i,Z_i)\right| \le C\delta_i + S_i^o(\tau,A_i)+2\Lambda_0^o(\tau)e^{\beta^o} = C < \infty.$$

Therefore, by Bernstein's inequality for independent bounded random variables we get, for each

$\varepsilon > 0$:

$$
\begin{aligned}
&P\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} H(t,A_i,Z_i)Q(t,A_i,Z_i)\right| > \varepsilon\right)\\
&\leq \exp\left(-\frac{\varepsilon^2/2}{C\varepsilon/\sqrt{n}+\mathrm{E}(\int_0^{\tau} H(t,A,Z)Q(t,A,Z))^2}\right)\\
&\leq \exp\left(-\frac{\varepsilon^2/2}{C\varepsilon/\sqrt{n}+\mathrm{E}(\sup_{[0,\tau],Z\in\mathcal{Z}} H^2(t,A,Z)\int_0^{\tau} Q^2(t,A,Z))}\right)\\
&= \exp\left(-\frac{\varepsilon^2/2}{C\varepsilon/\sqrt{n}+o_p(1)}\right) \to 0.
\end{aligned}
$$

□

Lemma 28. *For any $H(t)$ such that $\sup_{[0,\tau]}|H(t)| = o_p(1)$, under Assumptions 10-6 we have:*

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} H(t)\left\{w_i^* dN_i(t) + w_i^* G^o(t,A_i)dS_i^*(t,A_i)\right.\\
&\left.-\sum_{a=0,1} G^o(t,a)dS_i^*(t,a) - d\Lambda_0^o(t)\mathcal{S}^{(0)}(u;\beta^o,\pi^*,S^*,G^o)\right\}\\
&= o_p(n^{-1/2}),
\end{aligned}
$$

if either $\pi^* = \pi^o$ and $S^* = S^o$.

Proof of Lemma 28. We define:

$$
\begin{aligned}
Q(t,A_i,Z_i) &= w_i^* dM_i(t) + w_i^* G^o(t,A_i)dS_i^*(t,A_i) - \sum_{a=0,1} G^o(t,a)dS_i^*(t,a)\\
&\quad + d\Lambda_0^o(t)w_i^* G^o(t,A_i)S_i^*(t,A_i) - d\Lambda_0^o(t)\sum_{a=0,1} G^o(t,a)dS_i^*(t,a).
\end{aligned}
$$

By definition of $\mathcal{S}^{(0)}$, we want to prove that

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau H(t)Q(t,A_i,Z_i) = o_p(n^{-1/2}).$$

If $\pi^* = \pi^o$:

$$\begin{aligned}
&\mathrm{E}\left\{\int_0^\tau H(t)Q(t,A,Z)\right\} \\
&= \mathrm{E}\Bigg(\int_0^\tau \mathrm{E}\left\{\frac{A}{\pi^o(Z)}H(t)dM(t)|Z\right\} + \int_0^\tau \mathrm{E}\left\{\frac{1-A}{1-\pi^o(Z)}H(t)dM(t)|Z\right\} \\
&\quad + \mathrm{E}\left\{\frac{A}{\pi^o(Z)}H(t)G^o(t|A,Z)dS^*(t|A,Z)|Z\right\} + \mathrm{E}\left\{\frac{1-A}{1-\pi^o(Z)}H(t)G^o(t|A,Z)dS^*(t|A,Z)|Z\right\} \\
&\quad - \sum_{a=0,1} H(t)G^o(t|a,Z)dS^*(t|a,Z) \\
&\quad + d\Lambda_0^o(t)\Bigg[\mathrm{E}\left\{\frac{A}{\pi^o(Z)}H(t)G^o(t|A,Z)S^*(t|A,Z)|Z\right\} \\
&\quad + \mathrm{E}\left\{\frac{1-A}{1-\pi^o(Z)}H(t)G^o(t|A,Z)S^*(t|A,Z)|Z\right\} - \sum_{a=0,1} H(t)G^o(t|a,Z)S^*(t|a,Z)\Bigg]\Bigg).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathrm{E}\left\{\int_0^\tau H(t)Q(t,A,Z)\right\}\\
&=\mathrm{E}\left[\int_0^\tau \mathrm{E}\{H(t)dM(t)|A=1,Z\}+\int_0^\tau \mathrm{E}\{H(t)dM(t)|A=0,Z\}\right.\\
&\quad + H(t)G^o(t|1,Z)dS^*(t|1,Z)+H(t)G^o(t|0,Z)dS^*(t|0,Z)\\
&\quad - \sum_{a=0,1} H(t)G^o(t|a,Z)dS^*(t|a,Z)\\
&\quad +d\Lambda_0^o(t)\left\{H(t)G^o(t|A,Z)S^*(t|A,Z)\right.\\
&\quad \left.\left.+H(t)G^o(t|A,Z)S^*(t|A,Z)-\sum_{a=0,1} H(t)G^o(t|a,Z)S^*(t|a,Z)\right\}\right]\\
&=\mathrm{E}\left[\int_0^\tau H(t)dM^1(t)+\int_0^\tau H(t)dM^0(t)\right]=0.
\end{aligned}$$

On the other hand, if $S^*=S^o$, by (4.11) in the main document, we have:

$$\begin{aligned}
&\mathrm{E}\left\{\int_0^\tau H(t)Q(t,A,Z)\right\}\\
&=\mathrm{E}\left[\int_0^\tau \frac{A}{\pi^*(Z)}H(t)\mathrm{E}\{dM(t)|A,Z\}+\int_0^\tau \frac{1-A}{1-\pi^*(Z)}H(t)\mathrm{E}\{dM(t)|A,Z\}\right.\\
&\quad \int_0^\tau \frac{A}{\pi^*(Z)}H(t)\mathrm{E}\{dM(t)|A,Z\}-\int_0^\tau \frac{1-A}{1-\pi^*(Z)}H(t)\mathrm{E}\{dM(t)|A,Z\}\\
&\quad \left.+\sum_{a=0,1} H(t)\mathrm{E}\{dM(t)|a,Z\}\right]\\
&=\sum_{a=0,1}\mathrm{E}\left[\int_0^\tau H(t)dM^1(t)+\int_0^\tau H(t)dM^0(t)\right]=0.
\end{aligned}$$

The rest of the proof follows as the proof of Lemma 27. □

Lemma 29. *For any $H(t,Z)$ such that* $\sup_{Z\in\mathcal{Z}}|H(Z)| = o_p(1)$*, by Assumption 8 we have:*

$$\frac{1}{n}\sum_{i=1}^{n} H(Z_i)\left\{\frac{A_i}{\pi_i^o}e^{\beta^o A_i} - e^{\beta^o}\right\} = o_p(n^{-1/2}),$$

and

$$\frac{1}{n}\sum_{i=1}^{n} H(Z_i)\left\{\frac{A_i}{\pi_i^o} - 1\right\} = o_p(n^{-1/2}).$$

Proof of Lemma 29. We notice that

$$\mathrm{E}\left[H(Z)\left\{\frac{A}{\pi_i^o} - 1\right\}\right] = \mathrm{E}\left[H(Z)\mathrm{E}\left\{\frac{A}{\pi_i^o} - 1|Z\right\}\right] = 0,$$

and similarly

$$\mathrm{E}\left[H(Z)\left\{\frac{A}{\pi_i^o}e^{\beta^o A_i} - 1\right\}\right] = \mathrm{E}\left[H(Z)\mathrm{E}\left\{\frac{A}{\pi_i^o}e^{\beta^o A_i} - 1|Z\right\}\right] = 0,$$

The rest of the proof follows similarly to the proof of Lemma 27. □

Lemma 30. *For any $H(t,A,Z)$ such that* $\sup_{[0,\tau],Z\in\mathcal{Z}}|H(t,a,Z)| = o_p(1)$ *for* $a = 0,1$*, by Assumption 8 we have:*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{w_i^o H(t,A_i,Z_i) - \sum_{a=0,1} H(t,a,Z_i)\right\} = o_p(n^{-1/2}).$$

Proof of Lemma 30. We notice that

$$\begin{aligned}
&\mathrm{E}\left[\int_0^\tau \left\{ w^o H(t,A,Z) - \sum_{a=0,1} H(t,aZ) \right\}\right] \\
&= \mathrm{E}\left[\int_0^\tau \left\{ \frac{A}{\pi^o(Z)} H(t,1,Z) + \frac{1-A}{1-\pi^o(Z)} H(t,0,Z) - \sum_{a=0,1} H(t,a,Z) \right\}\right] \\
&= \mathrm{E}\left[\int_0^\tau \left\{ \frac{\mathrm{E}(A|Z)}{\pi^o(Z)} H(t,1,Z) + \frac{\mathrm{E}(1-A|Z)}{1-\pi^o(Z)} H(t,0,Z) - \sum_{a=0,1} H(t,a,Z) \right\}\right] = 0.
\end{aligned}$$

The rest of the proof follows similarly to the proof of Lemma 27. □

4.8.6 Simulation technique

We report here the specific steps used to generate simulated dataset.

- Simulate unobserved $V \sim F_v$ for some distribution F_v.
- Simulate covariates $Z \sim F_{z,v}$ for some distribution $F_{z,v}$ that depends on the unobserved v.
- Simulate treatment $A \sim B(\pi(Z))$ for some $\pi(Z)$.
- Define $u = F_v(V) \sim U(0,1)$. For $a = 0,1$ simulate potential $T(a)$ solving $\exp\left\{-e^{\beta a}\Lambda_0(t)\right\} = u$ for t for some chosen β and $\Lambda_0(t)$.
- Simulate $C(a) \sim F_{c,a}$ for some distribution $F_{c,a}$, for $a = 0,1$.
- Define $X(a) = \min\{T(a), C(a)\}$ and $\delta(a) = 1\{T(a) \leq C(a)\}$ for $a = 0,1$.
- Define the observed $X = X(a), \delta = \delta(a)$ for $a = A$.

4.8.7 HHP-HAAS dataset

Table 4.7: Summary of the HHP-HAAS data. Presented are mean (standard deviation) for the continuous variables, and frequency (%) for the cathegorical variables.

	Light Drinkers ($n = 1509$)	Heavy Drinkers ($n = 552$)
SystolicBP	148.88 (21.50)	150.76 (22.09)
Age	77.36 (4.06)	77.49 (4.10)
Education (in years)	11.02 (3.19)	10.17 (3.01)
ApoE genotype (yes)	278 (18.4%)	121 (21.9%)
HeartRate (in 30 secs)	31.31 (4.64)	31.90 (4.85)

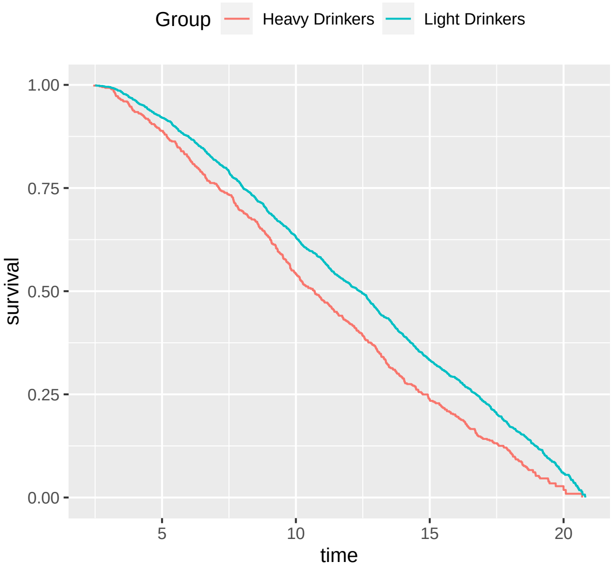


Figure 4.2: Kaplan-Meier curves for Light and Heavy drinkers for the HHP-HAAS data.

4.9 Acknowledgements

This research was supported by a grant from the National Institute on Aging; NIH/NIA R03 AG062432. Chapter 4, is currently being prepared for submission for publication of the material.

Rava, Denise; Bradic, Jelena; Xu, Ronghui. The dissertation author was the primary investigator and author of this material.

Bibliography

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. In *Lecture Notes in Statistics - 2: Mathematical Statistics and Probability Theory*, pages 1–25.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925.

Aalen, O. O., Stensrud, M. J., Didelez, V., Daniel, R., Røysland, K., and Strohmaier, S. (2020). Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biometrical Journal*, 62(3):532–549.

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

Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24:3927–3944.

Ash, A. and Shwartz, M. (1999). R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine*, 18:375–384.

Bai, X., Tsiatis, A. A., Lu, W., and Song, R. (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Analysis*, 23(4):585–604.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Benichou, J. and Gail, M. H. (1990). Estimates of absolute cause-specific risk in cohort studies. *Biometrics*, 46(3):813–826.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report, Univ. California, Berkeley*.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: some questions and an answer. *Statistica Sinica*, 11:863–886.

Biganzoli, E., Boracchi, P., Mariani, L., and Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17(10):1169–1186.

Blomberg, A., Wang, Y., Di, Q., Dominici, F., Schwartz, J., et al. (2019). Long-term effect of air pollution on hospital admissions among medicare participants using a doubly robust additive hazards model (drahm). *Environmental Epidemiology*, 3:355.

Bonsel, G., van't Veer, F., Habbema, J., Klompmaker, I., and Slooff, M. (1990). Use of prognostic models for assessment of value of liver transplantation in primary biliary cirrhosis. *the lancet*, 335(8688):493–497.

Brown, S. F., Branford, A. J., and Moran, W. (1997). On the use of artificial neural networks for the analysis of survival data. *IEEE Transactions on Neural Networks*, 8(5):1071–1077.

Brueckner, M., Titman, A., and Jaki, T. (2019). Instrumental variable estimation in semi-parametric additive hazards models. *Biometrics*, 75(1):110–120.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Lau, B., Adimora, A. A., and Study, W. I. H. (2014). Worth the weight: using inverse probability weighted cox models in aids research. *AIDS Research and Human Retroviruses*, 30(12):1170–1177.

Cefalu, M., Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., and Burgette, L. (2021). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups (version 2.3)*. CRAN R package.

Chan, P. H., Xu, R., and Chambers, C. D. (2018). A study of r^2 measure under the accelerated failure time models. *Communications in Statistics - Simulation and Computation*, 47:380–391.

Chauvel, C. and O'Quigley, J. (2017). Survival model construction guided by fit and predictive strength. *Biometrics*, 73(2):483–494.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038.

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4).

Christensen, E., Schlichting, P., Andersen, P. K., Fauerholdt, L., Schou, G., Pedersen, B. V., Juhl, E., Poulsen, H., Tygstrup, N., and for Liver Diseases, C. S. G. (1986). Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scandinavian Journal of Gastroenterology*, 21(2):163–174.

Cole, S. R., Hernán, M. A., Robins, J. M., Anastos, K., Chmiel, J., Detels, R., Ervin, C., Feldman, J., Greenblatt, R., Kingsley, L., et al. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, 158(7):687–694.

Comte, F., Gaïffas, S., and Guilloux, A. (2011). Adaptive estimation of the conditional intensity of marker-dependent counting processes. 47(4):1171–1196.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Cox, D. (1975). Partial likelihood. *Biometrika*, 62:269–276.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2020). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv preprint arXiv:2001.09887*.

Cui, Y., Zhu, R., Zhou, M., and Kosorok, M. (2017). Consistency of survival tree and forest models: splitting bias and correction. *Statistica Sinica (preprint)*.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics*, 17(3):1157–1167.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*, 10(1):1–7.

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

Dukes, O., Martinussen, T., Tchetgen Tchetgen, E. J., and Vansteelandt, S. (2019b). On doubly robust estimation of the hazard difference. *Biometrics*, 75(1):100–109.

Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82.

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

Feldman, H. I., Joffe, M., Robinson, B., Knauss, J., Cizman, B., Guo, W., Franklin-Becker, E., and Faich, G. (2004). Administration of parenteral iron and mortality among hemodialysis patients. *Journal of the American Society of Nephrology*, 15(6):1623–1632.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of Public Health*, 20(1):145–157.

Flander, P. and O'Quigley, J. (2019). Comparing kaplan-meier curves with delayed treatment effects: applications in immunotherapy trials. *Journal of the Royal Statistical Society, Series C*, 68:915–939.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Fosen, J., Ferkingstad, E., Borgan, Ø., and Aalen, O. O. (2006). Dynamic path analysis-a new approach to analyzing time-dependent covariates. *Lifetime data analysis*, 12(2):143–167.

Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*.

Freireich, E. J., Gehan, E., Frei, E., Schroeder, L., Wolman, I., Anbari, R., Burgert, E., Millis, S., Pinkel, D., Selawry, O., Moon, J., Gendel, B., Spurr, C., Storrs, R., Haurani, F., Hoogstraten, B., and Lee, S. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia. *Blood*, 21:699–716.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.

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

Gensheimer, M. F. and Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257.

Grambsch, P. M., Dickson, E. R., Wiesner, R. H., and Langworthy, A. (1989). Application of the mayo primary biliary cirrhosis survival model to mayo liver transplant patients. In *Mayo Clinic Proceedings*, volume 64, pages 699–704. Elsevier.

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

Grisan, E., Zandona, A., and Di Camillo, B. (2019). Deep convolutional neural network for survival estimation of amyotrophic lateral sclerosis patients.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Havercroft, W. and Didelez, V. (2012). Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine*, 31(30):4190–4206.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448.

Hernán, M. A., Lanoy, E., Costagliola, D., and Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology*, 98(3):237–242.

Hielscher, T., Zucknick, M., Werft, W., and Benner, A. (2010). On the prognostic value of survival models with application to gene expression signatures. *Statistics in Medicine*, 29(7-8):818–829.

Holt, J. (1978). Competing risk analyses with special reference to matched pair experiments. *Biometrika*, 65(1):159–165.

Honerkamp-Smith, G. and Xu, R. (2016). Three measures of explained variation for correlated survival data under the proportional hazards mixed-effects model. *Statistics in Medicine*, 35(23):4153–4165.

Hosmer, D. W., Lemeshow, S., and May, S. (2001). Applied survival analysis: Regression modeling of time to event data.

Hou, J., Bradic, J., and Xu, R. (2021). Treatment effect estimation under additive hazards models with high-dimensional confounding. *Journal of the American Statistical Association*, 116:1–42.

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

Hubbard, A. E., Van Der Laan, M. J., and Robins, J. M. (2000). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, volume 11, pages 135–177. Springer.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics*, 15(4):757–773.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3):841–860.

Jeffrey, G., Hoffman, N., and Reed, W. (1990). Validation of prognostic models in primary biliary cirrhosis. *Australian and New Zealand journal of medicine*, 20(2):107–110.

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

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data, 2nd Edition*. John Wiley & Sons, New York.

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

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24.

Kent, J. T. and O'Quigley, J. (1988). Measure of dependence for censored survival data. *Biometrika*, 75:525–534.

Kim, J. and Lee, S. (1998). Two-sample goodness-of-fit tests for additive risk models with censored observations. *Biometrika*, 85:593–603.

Klion, F. M., Fabry, T. L., Palmer, M., and Schaffner, F. (1992). Prediction of survival of patients with primary biliary cirrhosis: examination of the mayo clinic model on a group of patients with known endpoint. *Gastroenterology*, 102(1):310–313.

Kong, E., Xia, Y., and Zhong, W. (2019). Composite coefficient of determination and its application in ultrahigh dimensional variable screening. *Journal of the American Statistical Association*, 114(528):1740–1751.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995a). Hazard regression. *Journal of the American Statistical Association*, 90(429):78–94.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995b). The L2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics*, pages 143–157.

Kvamme, H. and Borgan, Ø. (2019). Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30.

Lange, T. and Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology*, 22:575–581.

Lee, C., Yoon, J., and Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, J., Fine, J., and Brookhart, A. (2015). Instrumental variable additive hazards models. *Biometrics*, 71(1):122–130.

Liao, L. and Ahn, H.-i. (2016). Combining deep learning and survival analysis for asset health management. *International Journal of Prognostic and Health Management*, 7:020.

Liestbl, K., Andersen, P. K., and Andersen, U. (1994). Survival analysis and neural nets. *Statistics in Medicine*, 13(12):1189–1200.

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

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

Luck, M., Sylvain, T., Cardinal, H., Lodi, A., and Bengio, Y. (2017). Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245*.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Markus, B. H., Dickson, E. R., Grambsch, P. M., Fleming, T. R., Mazzaferro, V., Klintmalm, G. B. G., Wiesner, R. H., Van Thiel, D. H., and Starzl, T. E. (1989). Efficacy of liver transplantation in patients with primary biliary cirrhosis. *New England Journal of Medicine*, 320(26):1709–1713.

Martinussen, T. (2010). Dynamic path analysis for event time data: large sample properties and inference. *Lifetime data analysis*, 16(1):85–101.

Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, 19:279–296.

Martinussen, T., Vansteelandt, S., Gerster, M., and Hjelmborg, J. v. B. (2011). Estimation of direct effects for survival data by using the aalen additive hazards model. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):773–788.

Müller, M., Döring, A., Küchenhoff, H., Lamina, C., Malzahn, D., Bickeböller, H., Vollmert, C., Klopp, N., Meisinger, C., Heinrich, J., et al. (2008). Quantifying the contribution of genetic variants for survival phenotypes. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(6):574–585.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and their Applications*, 39(1):153–180.

Nagpal, C., Li, X., and Dubrawski, A. (2020). Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *arXiv preprint arXiv:2003.01176*.

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

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

O'Quigley, J. (2008). *Proportional Hazards Regression*. Springer, New York.

O'Quigley, J. and Flandre, P. (1994). Predictive capability of proportional hazards regression. *Proceedings of the National Academy of Science USA*, 91:2310–2314.

O'Quigley, J. and Xu, R. (2012). *Handbook of Statistics in Clinical Oncology (3rd Ed.)*, chapter Explained variation and explained randomness for proportional hazards models, pages 487–503. Taylor & Francis Group, LLC.

O'Quigley, J., Xu, R., and Stare, J. (2005). Explained randomness in proportional hazards models. *Statistics in Medicine*, 24:479–489.

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and Van Der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2):147–185.

Preseley, A., Tilahun, A., Alonso, A., and Molenberghs, G. (2011). An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime Data Analysis*, 17:195–214.

Rebolledo, R. (1978). Sur les applications de la théorie des martingales à l'étude statistique d'une famille de processus ponctuels. In *Journées de Statistique des Processus Stochastiques*, pages 27–70. Springer.

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., and Yu, Y. (2019). Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4798–4805.

Reynaud-Bouret, P. et al. (2006). Penalized projection estimators of the aalen multiplicative intensity. *Bernoulli*, 12(4):633–661.

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

Robins, J. (1998). Marginal structural models. *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science*, pages 1–10.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Robins, J. M. and Ritov, Y. (1997). Towards a curse of dimensionality appropriate (CODA) asympototic theory for semiparametric models. *Statistics in Medicine*, 16:285–319.

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

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

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rotnitzky, A. and Robins, J. (2005). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics*, 4:2619–2625.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634.

Scheike, T. H. and Martinussen, T. (2006). *Dynamic Regression models for survival data*. Springer, NY.

Schemper, M. and Kaider, A. (1997). A new approach to estimate correlation coefficients in the presence of censoring and proportional hazards. *Computational Statistics and Data Analysis.*, 23:467–476.

Schumacher, M., Bastert, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., and Rauschecker, H. (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology*, 12(10):2086–2093.

Shen, Y. and Cheng, S. (1999). Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics*, 55(4):1093–1100.

Sjölander, A. and Vansteelandt, S. (2017). Doubly robust estimation of attributable fractions in survival analysis. *Statistical Methods in Medical Research*, 26(2):948–969.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2020). Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, 115:1–9.

Sterne, J. A., Hernán, M. A., Ledergerber, B., Tilling, K., Weber, R., Sendi, P., Rickenbach, M., Robins, J. M., Egger, M., Study, S. H. C., et al. (2005). Long-term effectiveness of potent antiretroviral therapy in preventing aids and death: a prospective cohort study. *The Lancet*, 366(9483):378–384.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053.

Tan, Z. (2019). On doubly robust estimation for logistic partially linear models. *Statistics & Probability Letters*, 155:108577.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology*, 26(3):402.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Tsiatis, A. A. (2006). Double-robust estimator of the average causal treatment effect. *Semiparametric Theory and Missing Data*, pages 323–337.

Van Der Laan, M. J., Laan, M., and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology*, 22(4):582.

VanderWeele, T. J. (2013). Unmeasured confounding and hazard scales: sensitivity analysis for total, direct, and indirect effects. *European Journal of Epidemiology*, 28(2):113–117.

Wang, C. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, 57(2):414–419.

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

Wei, G. (2008). *Semiparametric Methods for Estimating Cumulative Treatment Effects in the Presence of Non-proportional Hazards and Dependent Censoring.* PhD thesis, University of Michigan.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879.

Wellner, J. et al. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Wongvibulsin, S., Wu, K. C., and Zeger, S. L. (2020). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (rf-slam) data analysis. *BMC Medical Research Methodology*, 20(1):1–14.

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, 22:3527–3541.

Xu, R. and O'Quigley, J. (1999a). A R^2 measure of dependence for proportional hazards models. *Nonparametric Statistics*, 12:83–107.

Xu, R. and O'Quigley, J. (1999b). A R^2 type measure of dependence for proportional hazards models. *Nonparametric Statistics*, 12:83–107.

Yang, S., Pieper, K., and Cools, F. (2020). Semiparametric estimation of structural failure time models in continuous-time processes. *Biometrika*, 107(1):123–136.

Ying, A., Xu, R., and Murphy, J. (2019). Two-stage residual inclusion for survival data and competing risk- an instrumental variable approach with application to SEER-Medicare linked data. *Statistics in Medicine*, 38(1):125–138.

Yu, Z. and Van Der Laan, M. (2006). Double robust estimation in longitudinal marginal structural models. *Journal of Statistical Planning and Inference*, 136(3):1061–1089.

Yuen, K. C. and Burke, M. D. (1997). A test of fit for a semiparametric additive risk model. *Biometrika*, 84:631–639.

Zeng, D. and Chen, Q. (2010). Adjustment for missingness using auxiliary information in semiparametric regression. *Biometrics*, 66(1):115–122.

Zhang, M. and Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*, 67(3):740–749.

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

Zhang, M. and Schaubel, D. E. (2012b). Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics*, 68(4):999–1009.

Zhao, L. and Feng, D. (2019). Dnnsurv: Deep neural networks for survival analysis using pseudo values. *arXiv preprint arXiv:1908.02337*.

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

Zheng, C., Dai, R., Hari, P. N., and Zhang, M.-J. (2017). Instrumental variable with competing risk model. *Statistics in Medicine*, 36:1240–1255.

Zheng, W., Petersen, M., and Van Der Laan, M. J. (2016). Doubly robust and efficient estimation of marginal structural models for the hazard function. *The International Journal of Biostatistics*, 12(1):233–252.

Zhong, C. and Tibshirani, R. (2019). Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171*.

Zhu, X., Yao, J., and Huang, J. (2016). Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE.