Biometrika (2020), **107**, 1, *pp.* 123–136
Printed in Great Britain
doi: 10.1093/biomet/asz057
Advance Access publication 29 October 2019

Semiparametric estimation of structural failure time models in continuous-time processes

BY S. YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.

syang24@ncsu.edu

K. PIEPER

Duke Clinical Research Institute, Duke University, 300 W. Morgan Street, Durham, North Carolina 27705, U.S.A.

karen.pieper@duke.edu

AND F. COOLS

Department of Cardiology, AZ Klina, Augustijnslei 100, 2930 Brasschaat, Belgium

frank.cools@klina.be

SUMMARY

Structural failure time models are causal models for estimating the effect of time-varying treatments on a survival outcome. G-estimation and artificial censoring have been proposed for estimating the model parameters in the presence of time-dependent confounding and administrative censoring. However, most existing methods require manually pre-processing data into regularly spaced data, which may invalidate the subsequent causal analysis. Moreover, the computation and inference are challenging due to the nonsmoothness of artificial censoring. We propose a class of continuous-time structural failure time models that respects the continuous-time nature of the underlying data processes. Under a martingale condition of no unmeasured confounding, we show that the model parameters are identifiable from a potentially infinite number of estimating equations. Using the semiparametric efficiency theory, we derive the first semiparametric doubly robust estimators, which are consistent if the model for the treatment process or the failure time model, but not necessarily both, is correctly specified. Moreover, we propose using inverse probability of censoring weighting to deal with dependent censoring. In contrast to artificial censoring, our weighting strategy does not introduce nonsmoothness in estimation and ensures that resampling methods can be used for inference.

Some key words: Causality; Cox proportional hazards model; Discretization; Observational study; Semiparametric analysis; Survival data.

1. INTRODUCTION

Confounding by indication is common in observational studies and obscures the causal relationship between the treatment and outcome, (Robins et al., 1992). In longitudinal observational studies, this phenomenon becomes more pronounced because of time-varying confounding, when

there are time-dependent covariates that predict the subsequent treatment and outcome, and are also affected by the past treatment history. In this case standard regression methods, whether adjusting for confounders or not, are fallible (Robins et al., 2000; Daniel et al., 2013).

Structural failure time models (Robins & Tsiatis, 1991; Robins, 1992) and marginal structural models (Robins, 2000; Hernán et al., 2001) have been used to handle time-varying confounding effectively. Structural failure time models simulate the potential failure time outcome that would have been observed in the absence of treatment, referred to as the potential baseline failure time, by removing the treatment effect, while marginal structural models specify the marginal relationship between potential outcomes under different treatments, possibly adjusting for the baseline covariates. Structural failure time models have certain features that are more desirable than marginal structural models (Robins, 2000): structural failure time models allow for the modelling of time-varying treatment modification effects using the post-baseline time-dependent covariates; they are more flexible in terms of translating biological hypotheses into their parameters (Robins, 1998b; Lok, 2008); and g-estimation (Robins, 1998b) for structural failure time models does not require the probability of receiving treatment at each time-point to be positive for all subjects.

Most structural failure time models specify deterministic relationships between the observed failure time and the potential baseline failure time, and are therefore rank preserving (see, e.g., Mark & Robins, 1993a,b; Robins & Greenland, 1994; Robins, 2002; Hernán et al., 2005). Moreover, existing g-estimation approaches often use a discrete-time set-up, which requires all subjects to be followed at the same prefixed time-points. However, in practical situations, the variables and processes are more likely to be measured at irregularly spaced time-points, which may not be the same for all subjects (Robins, 1998a). To use existing estimators, one needs to discretize the timeline and recreate the measurements at each time-point, for example by averaging observations within the given time-point or by imputation if there are no observations. Such data pre-processing may distort the relationship between variables and cast doubt on the sequential randomization assumption, which is essential to justification of the discrete-time g-estimation (Zhang et al., 2011). In the literature, much less work has addressed non-rank-preserving continuous-time causal models; exceptions include Robins (1998b), Lok et al. (2004) and Lok (2008, 2017). Robins (1998b) conjectured that g-estimation extends to settings with continuous-time processes, but still relies on the rank-preserving assumption. Recently, Lok (2017) presented a formal proof of the extension conjecture without the assumption of rank preservation.

Despite these advances, estimation for continuous-time structural failure time models is largely underdeveloped. Existing g-estimation is singly robust, in the sense that it relies on a correct model specification for the treatment process. In the literature of missing data analysis and causal inference, many authors have proposed doubly robust estimators that require either one of the two model components to be correctly specified (Robins et al., 1994; Scharfstein et al., 1999; Van Der Laan et al., 2002; Lunceford & Davidian, 2004; Bang & Robins, 2005; Robins et al., 2007; Cao et al., 2009; Lok & DeGruttola, 2012). Yang & Lok (2016) constructed a doubly robust test procedure for structural nested mean models. To the best of our knowledge, a doubly robust estimator for structural failure time models does not exist.

We develop a general framework for structural failure time models with continuous-time processes. We relax the local rank-preservation condition by specifying a distributional rather than deterministic relationship between the treatment process and the potential baseline failure time. We impose a martingale condition of no unmeasured confounding, which serves as the basis for identification and estimation. Under the semiparametric model characterized by the structural failure time model and the no unmeasured confounding assumption, we develop a class of regular asymptotically linear estimators. This class of estimators contains the semiparametric efficient

estimators (Bickel et al., 1993; Tsiatis, 2006). We further construct an optimal member among a wide class of semiparametric estimators that are relatively simple to compute. Moreover, we show that our estimators are doubly robust in the sense that they are consistent if either the model for the treatment process is correctly specified or the failure time model is correctly specified, but not necessarily both. Our framework is readily applicable to the traditional discrete-time settings.

In the presence of censoring, Robins and coauthors have introduced the notion of the potential censoring time and proposed a way of using this information to estimate the treatment effect. This approach may artificially terminate follow-up for some subjects before their observed failure or censoring times, so it is often called artificial censoring. It works only for administrative censoring when follow-up ends at a prespecified date, and it fails to provide consistent estimators for dependent censoring (Rotnitzky & Robins, 1995), which likely occurs due to drop-out of subjects. Moreover, the computation and inference are challenging because of the nonsmoothness of artificial censoring (Joffe, 2001; Joffe et al., 2012). To overcome these limitations, we propose using inverse probability of censoring weighting. In contrast to artificial censoring, our weighting strategy is smooth and ensures that resampling methods can be used for inference, which is straightforward to implement in practice.

2. Notation, models and assumptions

2.1. *Notation*

We assume that n subjects constitute a random sample from a larger population of interest and are therefore independent and identically distributed. For notational simplicity, we suppress the subscript i for subjects. Let T be the observed failure time. Let L_t be a multi-dimensional covariate process, and let A_t be the binary treatment process, i.e., $A_t = 1$ if the subject is on treatment at time t and $A_t = 0$ if the subject is off treatment at time t. We assume that all subjects received treatment at baseline and may discontinue treatment during follow-up. We also assume that treatment discontinuation is permanent, i.e., if $A_t = 0$ then $A_u = 0$ for all $u \geqslant t$. Let V be the time to treatment discontinuation or failure, whichever comes first, and let Γ be the binary indicator of treatment discontinuation at time V. To ensure regularity, we assume that all continuous-time processes are càdlàg, i.e., the processes are continuous from the right and have limits from the left. Let $H_t = (L_t, A_{t-})$ be the combined covariates and treatment process, where A_{t-} denotes the treatment just before time t. We use an overbar to denote the history; for example, $\bar{H}_t = (H_u : 0 \leqslant u \leqslant t)$ is the history of the covariates and treatment process until time t. Following Cox & Oakes (1984), we assume that there exists a potential baseline failure time U, representing the failure time had the treatment always been withheld. The full data consist of $F = (T, \bar{H}_T)$. Up until § 4 we assume that there is no censoring before T.

2.2. *Structural failure time model*

The structural failure time model specifies the relationship between the potential baseline failure time U and the actual observed failure time T. We assume that given any $\bar{H}_t$,

$$U \sim U(\psi^*) = \int_0^T \exp\big[\{\psi_1^* + \psi_2^{*\mathrm{T}} g(L_u)\}A_u\big]\,\mathrm{d}u, \tag{1}$$

where $\sim$ means has the same distribution as and $\psi^{*\mathrm{T}} = (\psi_1^*, \psi_2^{*\mathrm{T}})$ is a p-vector of unknown parameters. Model (1) entails that the treatment effect is to accelerate or decelerate the failure

time relative to the potential baseline failure time U. Intuitively, $\exp[\{\psi_1^* + \psi_2^{*\mathrm{T}} g(L_t)\}A_t]$ can be interpreted as the effect rate of the treatment on the outcome, possibly modified by the time-varying covariate $g(L_t)$. To aid understanding of the model, consider a simplified model $U(\psi^*) = \int_0^T \exp(\psi_1^* A_u)\,\mathrm{d}u$. The multiplicative factor $\exp(\psi_1^*)$ describes the relative increase or decrease in the failure time had the subject continuously received treatment compared to had the treatment always been withheld.

Remark 1. The rank-preserving structural failure time model specifies a deterministic relationship instead of a distributional relationship between the failure times, i.e., it uses $=$ instead of $\sim$ in model (1). Then, for subjects i and j who have the same observed treatment and covariate history, $T_i < T_j$ must imply $U_i < U_j$. This may be restrictive in practice. In contrast, we link the distribution of the potential baseline failure time and the distribution of the actual failure time after removing the treatment effect. Specifically, we assume that the distributions of U and $U(\psi^*)$ are the same given past treatment and covariates, thus avoiding the rank-preserving restriction.

2.3. *No unmeasured confounding*

The model parameter ψ^* is not identifiable in general, because U is missing for all subjects. To identify and estimate ψ^*, we impose the following assumption (Yang et al., 2018).

Assumption 1 (No unmeasured confounding). The hazard of treatment discontinuation is

$$\begin{aligned}\lambda_V(t \mid F, U) &= \lim_{h\to 0} h^{-1}\mathrm{pr}(t \leqslant V < t+h, \Gamma = 1 \mid F, U, V \geqslant t)\\ &= \lim_{h\to 0} h^{-1}\mathrm{pr}(t \leqslant V < t+h, \Gamma = 1 \mid \bar{H}_t, V \geqslant t) = \lambda_V(t \mid \bar{H}_t). \qquad (2)\end{aligned}$$

Assumption 1 implies that $\lambda_V(t \mid F, U)$ depends only on the past treatment and covariate history up to time t, $\bar{H}_t$, but not on the future variables and U. This assumption holds if the set of historical covariates contains all prognostic factors for the failure time that affect the decision of discontinuing treatment at time t.

For an equivalent representation of the treatment process A_t, we define the counting process $N_V(t) = I(V \leqslant t, \Gamma = 1)$ and the at-risk process $Y_V(t) = I(V \geqslant t)$ (Andersen et al., 1993). Let $\sigma(H_t)$ be the σ-field generated by H_t, and let $\sigma(\bar{H}_t)$ be the σ-field generated by $\bigcup_{u\leqslant t}\sigma(H_u)$. We show in the Supplementary Material that under model (1), (2) implies that

$$\lambda_V\{t \mid \bar{H}_t, U(\psi^*)\} = \lambda_V(t \mid \bar{H}_t).$$

Thus, under common regularity conditions for the counting process, $M_V(t) = N_V(t) - \int_0^t \lambda_V(u \mid \bar{H}_u) Y_V(u)\,\mathrm{d}u$ is a martingale with respect to $\sigma\{U(\psi^*), \bar{H}_t\}$, which renders ψ^* identifiable.

3. SEMIPARAMETRIC ESTIMATION

We consider the semiparametric model characterized by (1) and Assumption 1. We derive a regular asymptotically linear estimator $\hat{\psi}$ of ψ^*, such that

$$n^{1/2}(\hat{\psi} - \psi^*) = P_n\Phi(F) + o_{\mathrm{p}}(1),$$

where P_n is the empirical measure induced by $F_1, \ldots, F_n$, i.e., $P_n\Phi(F) = n^{-1}\sum_{i=1}^n \Phi(F_i)$, and $\Phi(F)$ is the influence function of $\hat{\psi}$, which has zero mean and finite and nonsingular variance.

Let $f_F(T, \bar{H}_T; \psi, \theta)$ be the semiparametric likelihood function based on a single variable F, where ψ is the primary parameter of interest and θ is the infinite-dimensional nuisance parameter. A fundamental result of Bickel et al. (1993) states that the influence functions for regular asymptotically linear estimators lie in the orthogonal complement of the nuisance tangent space, denoted by $\Lambda^{\perp}$. We characterize $\Lambda^{\perp}$ in the following theorem, the proof of which is given in the Supplementary Material.

THEOREM 1. *Under model* (1) *and Assumption* 1*, the orthogonal complement of the nuisance tangent space for* ψ^* *is*

$$\Lambda^{\perp} = \left\{ \int_0^{\infty} \big(h_u\{U(\psi^*), \bar{H}_u\} - E\big[h_u\{U(\psi^*), \bar{H}_u\} \mid \bar{H}_u, V \geqslant u\big] \big) \, \mathrm{d}M_V(u) \right\}$$

for all p-dimensional $h_u\{U(\psi^*), \bar{H}_u\}$.

The score function of ψ^* is $S_\psi(F) = \partial \log f_F(T, \bar{H}_T; \psi, \theta)/\partial\psi$ evaluated at (ψ^*, θ^*). Following Bickel et al. (1993), the efficient score for ψ^* is $S_{\mathrm{eff}}(F) = \Pi\{S_\psi(F) \mid \Lambda^{\perp}\}$, where Π is the projection operator in the Hilbert space. The efficient influence function is $\Phi(F) = E\{S_{\mathrm{eff}}(F) S_{\mathrm{eff}}(F)^{\mathrm{T}}\}^{-1} S_{\mathrm{eff}}(F)$, with the variance $[E\{S_{\mathrm{eff}}(F) S_{\mathrm{eff}}(F)^{\mathrm{T}}\}]^{-1}$ achieving the semiparametric efficiency bound. However, the analytical form of $S_\psi(F)$ is intractable in general. To facilitate estimation, we focus on a reduced class of $\Lambda^{\perp}$ with $h_u\{U(\psi^*), \bar{H}_u\} = c(\bar{H}_u) U(\psi^*)$ for $c(\bar{H}_u) \in \mathbb{R}^p$, leading to the following estimating function for ψ^*:

$$G(\psi; F) = \int_0^{\infty} c(\bar{H}_u) \big[U(\psi) - E\{U(\psi) \mid \bar{H}_u, V \geqslant u\} \big] \, \mathrm{d}M_V(u). \tag{3}$$

Because of the no unmeasured confounding assumption, $U(\psi^*) \perp\!\!\!\perp M_V(u) \mid (\bar{H}_u, V \geqslant u)$ and so $E\{G(\psi^*; F)\} = 0$. We obtain the estimator of ψ^* by solving

$$P_n\{G(\psi; F)\} = 0. \tag{4}$$

Within this class, we show that the optimal choice of $c(\bar{H}_u)$ is

$$c^{\mathrm{opt}}(\bar{H}_u) = E\{\partial \dot{U}_u(\psi)/\partial\psi \mid \bar{H}_u, V = u\} \big[\mathrm{var}\{U(\psi) \mid \bar{H}_u, V \geqslant u\} \big]^{-1}. \tag{5}$$

In practice, we require working models to be posited for approximating $c^{\mathrm{opt}}(\bar{H}_u)$; see the example in the simulation study. Compared to naive choices, such as $c(\bar{H}_u) = \{A_u, A_u g(L_u)^{\mathrm{T}}\}^{\mathrm{T}}$ for model (1), our simulation results show that using the optimal choice yields gains in estimation efficiency.

In (4), we assume that the hazard function for the treatment process and $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ are known. In practice, they are often unknown and must be modelled and estimated from the data. We posit a proportional hazards model with time-dependent covariates,

$$\lambda_V(t \mid \bar{H}_t; \gamma_V) = \lambda_{V,0}(t) \exp\{\gamma_V^{\mathrm{T}} g_V(t, \bar{H}_t)\},$$

where $\lambda_{V,0}(t)$ is unknown and nonnegative, $g_V(t, \bar{H}_t)$ is a prespecified function of t and $\bar{H}_t$, and γ_V is a vector of unknown parameters. We also posit a working model $E\{U(\psi) \mid \bar{H}_u, V \geqslant u; \xi\}$ indexed by ξ. We show that the estimating equation for ψ^* achieves the double robustness or double protection (Rotnitzky & Vansteelandt, 2015).

THEOREM 2 (Double robustness). *Under model* (1) *and Assumption* 1*, the estimating equation* (4) *for* ψ^* *is unbiased if either the model for the treatment process is correctly specified or the failure time model* $E\{U(\psi) \mid \bar{H}_u, V \geqslant u; \xi\}$ *is correctly specified, but not necessarily both.*

4. CENSORING

4.1. *Inverse probability of censoring weighting*

In most studies, the failure time is subject to right censoring. We now introduce C, the time to censoring. The observed data are $O = \{X = \min(T, C), \Delta = 1(T \leqslant C), \bar{H}_X\}$. In the presence of censoring, we may not observe T, so it may not be feasible to solve the estimating equation (4). A naive solution is to replace T in $U(\psi)$ by X and use $\tilde{U}(\psi) = \int_0^X \exp(\psi A_s)\,\mathrm{d}s$; however, $\tilde{U}(\psi^*)$ depends on the whole treatment process and is therefore not independent of $M_V(t)$ given $(\bar{H}_t, V \geqslant t)$, which renders the estimating equation (4) biased (Hernán et al., 2005). Robins (1998b) proposed a strategy for dealing with administrative censoring, a censoring mechanism which occurs when subjects are censored due to the fact that the study ended at a known calendar date. In this case, C is independent of all other variables. In Robins's strategy, $U(\psi)$ is replaced by a function of $U(\psi)$ and C, which is always observable. For illustration, consider $U(\psi) = \int_0^T \exp(\psi A_u)\,\mathrm{d}u$ and

$$C(\psi) = \min_{a_s \in \{0,1\}} \int_0^C \exp(\psi a_s)\,\mathrm{d}s = \begin{cases} C, & \psi \geqslant 0, \\ C\exp(\psi), & \psi < 0. \end{cases}$$

Then $\tilde{U}(\psi^*) = \min\{U(\psi^*), C(\psi^*)\}$ and $\Delta(\psi^*) = 1\{U(\psi^*) < C(\psi^*)\}$ are two functions that are independent of $M_V(t)$ given $(\bar{H}_t, V \geqslant t)$ and always computable; see the Supplementary Material. The g-estimator is constructed based on $\tilde{U}(\psi)$ and $\Delta(\psi)$. In this approach, for subjects with $T < C$ it is possible that $U(\psi) > C(\psi)$ and $\Delta(\psi) = 0$, i.e., those subjects who actually were observed to fail are treated as if they were censored. Therefore, this approach is often called artificial censoring. Artificial censoring suffers from many drawbacks. First, the resulting estimating equation is not smooth in ψ, and therefore the estimation and inference are challenging (Joffe et al., 2012). Second, if the censoring mechanism is dependent, the estimators will be inconsistent (Robins, 1998b). To avoid the disadvantages of artificial censoring and also allow for more general censoring mechanisms, we consider using inverse probability of censoring weighting. Robins (1998b) suggested and Witteman et al. (1998) applied the weighting approach to deal with censoring by competing risks in deterministic structural failure time models with discretized data. We now assume an ignorable censoring mechanism as follows.

Assumption 2. The hazard of censoring is

$$\begin{aligned}\lambda_C(t \mid F, T > t) &= \lim_{h \to 0} h^{-1} \mathrm{pr}(t \leqslant C < t + h \mid C \geqslant t, F, T > t) \\ &= \lim_{h \to 0} h^{-1} \mathrm{pr}(t \leqslant C < t + h \mid C \geqslant t, \bar{H}_t, T > t) = \lambda_C(t \mid \bar{H}_t, T > t),\end{aligned}$$

written as $\lambda_C(t \mid \bar{H}_t)$ for short.

Assumption 2 says that $\lambda_C(t \mid F, T > t)$ depends only on the past treatment and covariate history up to time t, but not on the future variables and failure time. This assumption holds if the

set of historical covariates contains all prognostic factors for the failure time that affect the loss to follow-up at time t. Under this assumption, the missing data due to censoring are missing at random (Rubin, 1976). In the presence of censoring, V is redefined to be the time to treatment discontinuation, failure or censoring, whichever comes first. We show in the Supplementary Material that $\lambda_V(t \mid \bar{H}_t)$ is equal to $\lambda_V(t \mid \bar{H}_t, C \geqslant t)$ and so can be estimated conditional on $V \geqslant t$ with the new definition of V. From $\lambda_C(t \mid \bar{H}_t)$ we define $K_C(t \mid \bar{H}_t) = \exp\{-\int_0^t \lambda_C(u \mid \bar{H}_u)\,\mathrm{d}u\}$, which is the probability of the subject not being censored before time t. For regularity, we also impose a positivity condition on $K_C(t \mid \bar{H}_t)$.

Assumption 3 (Positivity). There exists a constant δ such that with probability 1, $K_C(t \mid \bar{H}_t) \geqslant \delta > 0$ for t in the support of T.

Under Assumptions 1–3, ψ^* is identifiable; see the Supplementary Material for a proof. Following Rotnitzky et al. (2009), the main idea of inverse probability of censoring weighting is to redistribute the weights for the censored subjects to the remaining uncensored subjects.

THEOREM 3. *Under Assumptions 1–3, the unbiased estimating equation for ψ^* is*

$$P_n\left\{\frac{\Delta}{K_C(T \mid \bar{H}_T)}\, G(\psi; F)\right\} = 0, \tag{6}$$

where $G(\psi; F)$ is as defined in (3).

Theorem 3 assumes that $\lambda_C(t \mid \bar{H}_t)$ is known. As was done for $\lambda_V(t \mid \bar{H}_t)$, we posit a proportional hazards model with time-dependent covariates,

$$\lambda_C(t \mid \bar{H}_t) = \lambda_{C,0}(t) \exp\{\gamma_C^{\mathrm{T}} g_C(t, \bar{H}_t)\},$$

where $\lambda_{C,0}(t)$ is unknown and nonnegative, $g_C(t, \bar{H}_t)$ is a prespecified function of t and $\bar{H}_t$, and γ_C is a vector of unknown parameters.

To summarize, the algorithm for developing an estimator of ψ^* is as follows.

Step 1. Using the data $(V_i, \Gamma_i, \bar{H}_{V_i,i})$ $(i = 1, \ldots, n)$, obtain estimators for $\lambda_V(t \mid \bar{H}_t) = \lambda_{V,0}(t) \exp\{\gamma_V^{\mathrm{T}} g_V(t, \bar{H}_t)\}$ and $M_V(t)$. To estimate γ_V, treat the treatment discontinuation as failure and the failure event and censoring as censored observations in the time-dependent proportional hazards model. Once we have an estimate of γ_V, $\hat{\gamma}_V$, we can estimate the cumulative baseline hazard, $\lambda_{V,0}(t)\,\mathrm{d}t$, using the Breslow estimator

$$\hat{\lambda}_{V,0}(t)\,\mathrm{d}t = \frac{\sum_{i=1}^n \mathrm{d}N_{V,i}(t)}{\sum_{i=1}^n \exp\{\hat{\gamma}_V^{\mathrm{T}} g_V(t, \bar{H}_{t,i})\} Y_{V_i}(t)}.$$

Then we obtain $\hat{M}_V(t) = N_V(t) - \int_0^t \exp\{\hat{\gamma}_V^{\mathrm{T}} g_V(u, \bar{H}_u)\} \hat{\lambda}_{V,0}(u) Y_V(u)\,\mathrm{d}u$.

Step 2. Using the data $(X_i, \Delta_i, \bar{H}_{X_i,i})$ $(i = 1, \ldots, n)$, obtain estimators for $\lambda_C(t \mid \bar{H}_t) = \lambda_{C,0}(t) \exp\{\gamma_C^{\mathrm{T}} g_C(t, \bar{H}_t)\}$ and $K_C(T_i \mid \bar{H}_{T_i})$. To estimate γ_C, treat censoring as failure and the failure event as censored observations in the time-dependent proportional hazards model. Once we have an estimate of γ_C, $\hat{\gamma}_C$, we can estimate $\lambda_{C,0}(t)\,\mathrm{d}t$ using the Breslow estimator

$$\hat{\lambda}_{C,0}(t)\,\mathrm{d}t = \frac{\sum_{i=1}^n \mathrm{d}N_{C,i}(t)}{\sum_{i=1}^n \exp\{\hat{\gamma}_C^{\mathrm{T}} g_C(t, \bar{H}_{t,i})\} Y_{C_i}(t)},$$

where $N_C(t) = I(C \leqslant t, \Delta = 0)$ and $Y_C(t) = I(C \geqslant t)$ are the counting process and the at-risk process of observing censoring, respectively. Then we estimate $K_C(t \mid \bar{H}_t)$ by

$$\hat{K}_C(t \mid \bar{H}_t) = \prod_{0 \leqslant u \leqslant t} \big[1 - \exp\{\hat{\gamma}_C^{\mathrm{T}} g_C(u, \bar{H}_u)\} \hat{\lambda}_{C,0}(u)\, \mathrm{d}u\big].$$

Step 3. We obtain the estimator $\hat{\psi}$ of ψ by solving

$$P_n \left\{ \frac{\Delta}{\hat{K}_C(T \mid \bar{H}_T)} \int c(\bar{H}_u) \big[U(\psi) - E\{U(\psi) \mid \bar{H}_u, V \geqslant u; \hat{\xi}\} \big]\, \mathrm{d}\hat{M}_V(u) \right\} = 0, \tag{7}$$

where we estimate $E\{U(\psi) \mid \bar{H}_u, V \geqslant u; \xi\}$ by regressing $\hat{K}_C(T \mid \bar{H}_T)^{-1} \Delta U(\psi)$ on (X_0, L_u, u) restricted to subjects with $V \geqslant u$. The estimating equation (7) is continuously differentiable in ψ, and hence can generally be solved using a Newton–Raphson procedure (Atkinson, 1989). For example, one can use the `multiroot` function in `R` (R Development Core Team, 2020).

Remark 2. It is worth discussing the connection between the proposed framework and the existing framework for the discrete-time setting. If the processes take observations at discrete times $\{t_0, \ldots, t_K\}$, then for $t = t_m$, $\bar{H}_t = \{H_{t_1}, \ldots, H_{t_m}\}$, $\mathrm{d}N_T(t)$ is a binary treatment indicator, and $\int_0^t \lambda_T(u \mid \bar{H}_u) Y_T(u)\, \mathrm{d}u$ becomes the propensity score $\mathrm{pr}\{\mathrm{d}N_T(t) = 1 \mid \bar{H}_t\}$. As a result, in the special case where $E\{U(\psi) \mid \bar{H}_u, V \geqslant u; \hat{\xi}\}$ is zero, (7) simplifies to the existing estimating equation for ψ^*. Importantly, (7) provides, for the first time in the literature, a semiparametric doubly robust estimator $\hat{\psi}$ even for the discrete-time setting, in the sense that $\hat{\psi}$ is consistent if either the model for the treatment process or the failure time model is correctly specified, under correct model specifications for the treatment effect mechanism and the censoring mechanism.

4.2. *Asymptotic theory and variance estimation*

In this section we discuss the asymptotic properties of our proposed estimator; the technical details are presented in the Supplementary Material. To reflect the dependence of the estimating equation on the nuisance models, write (7) as $P_n \Phi(\psi, \hat{\xi}, \hat{M}_V, \hat{K}_C; F) = 0$, where

$$\begin{aligned} \Phi(\psi, \xi, M_V, K_C; F) = {} & \{K_C(T \mid \bar{H}_T)\}^{-1} \Delta \\ & \times \int c(\bar{H}_u) \big[U(\psi) - E\{U(\psi) \mid \bar{H}_u, V \geqslant u; \xi\} \big]\, \mathrm{d}M_V(u). \end{aligned}$$

Let the probability limits of $\hat{\xi}$, $\hat{M}_V$ and $\hat{K}_C$ be ξ^*, M_V^* and K_C^*, respectively. We impose standard regularity conditions for Z-estimators (van der Vaart & Wellner, 1996). Roughly speaking, these conditions restrict the flexibility and convergence rates of the nuisance estimators; for example, we assume that $\Phi(\psi, \xi, M_V, K_C; F)$ and $\partial \Phi(\psi, \xi, M_V, K_C; F)/\partial \psi$ belong to P-Donsker classes. The regularity conditions ensure that

$$\begin{aligned} E\Bigg(\int c(\bar{H}_u) \Bigg[E\left\{ \begin{pmatrix} U(\psi^*) \\ \partial U(\psi^*)/\partial \psi \end{pmatrix} \,\middle|\, \bar{H}_u, V \geqslant u; \hat{\xi} \right\} \\ - E\left\{ \begin{pmatrix} U(\psi^*) \\ \partial U(\psi^*)/\partial \psi \end{pmatrix} \,\middle|\, \bar{H}_u, V \geqslant u; \xi^* \right\} \Bigg] \mathrm{d}\{\hat{M}_V(u) - M_V^*(u)\} \Bigg) = o(n^{-1/2}). \end{aligned}$$

Under Assumptions 3 and further assumptions in the Supplementary Material if K_C is correctly specified and if either $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ or M_V is correctly specified, $\hat{\psi}$ solving (7) with

the estimated nuisance models is still consistent and asymptotically normal, with the influence function $\tilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F)$.

We can estimate the variance of $\hat{\psi}$ either by the empirical variance of the estimated influence function or by resampling. If all the nuisance models, ξ, M_V and K_C, are correctly specified, we obtain an analytical expression for $\tilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F)$. We can then estimate $\tilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F)$ by plugging in estimates of $\psi^*, \xi^*, M_V^*, K_C^*$ and the required expectations, denoted by $\hat{\Phi}(\hat{\psi}, \hat{\xi}, \hat{M}_V, \hat{K}_C; F)$. Then the estimated variance of $n^{1/2}(\hat{\psi} - \psi^*)$ is

$$P_n\{\hat{\Phi}(\hat{\psi}, \hat{\xi}, \hat{M}_V, \hat{K}_C; F)\hat{\Phi}(\hat{\psi}, \hat{\xi}, \hat{M}_V, \hat{K}_C; F)^{\mathrm{T}}\}. \tag{8}$$

However, when one of ξ and M_V is correctly specified, but not both, characterizing $\tilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F)$ is difficult, and hence approximating (8) is no longer feasible. To avoid this technical difficulty, we recommend estimating the asymptotic variance by resampling methods such as the bootstrap and jackknife (Efron, 1979; Efron & Stein, 1981). In this case, the resampling works because $\hat{\psi}$ is regular and asymptotically normal.

5. Simulation study

We evaluate the finite-sample performance of the proposed estimator on simulated datasets. We generate U from Ex(0.2) and generate the covariate process (X_0, L_t) had the treatment always been withheld. We generate X_0 from Ber(0.55). To generate L_t, we first generate a 1×3 row vector following a multivariate normal distribution with mean $0.2U - 4$ and covariance $0.7^{|i-j|}$ for $i, j = 1, 2, 3$. This vector represents the values of L_t at times $t_1 = 0$, $t_2 = 5$ and $t_3 = 10$. We assume that the time-dependent variable remains constant between measurements. We generate the time until treatment discontinuation, V_1, according to a proportional hazards model $\lambda_V(t \mid X_0, \bar{L}_t) = 0.15\exp(0.15X_0 + 0.15L_t)$. This determines the treatment process A_t, i.e., $A_t = 1$ if $t \leqslant V_1$ and $A_t = 0$ if $t > V_1$. The observed time-dependent covariate process is L_t if $t \leqslant V_1$ and $L_t + \log(t - V_1)$ if $t > V_1$, to reflect the fact that the covariate process is affected after treatment discontinuation. Let the history of covariates and treatment up to time t be $\bar{H}_t = (X_0, \bar{L}_t, \bar{A}_{t-})$. We generate T according to $U \sim \int_0^T \exp(\psi^* A_u)\,\mathrm{d}u$ as follows. Let $T_1 = U\exp(-\psi^*)$. If $T_1 < V_1$, then $T = T_1$; otherwise $T = U + V_1 - V_1\exp(\psi^*)$. Under the above data-generating mechanism, the potential failure time under $\bar{a}_T$ also follows a Cox marginal structural model with the hazard rate at u, $\lambda_0(u)\exp(\psi^* A_u)$ (Young et al., 2010). We generate C according to a proportional hazards model with $\lambda_C(t \mid X_0, \bar{L}_t, C \geqslant t) = 0.025\exp(0.15X_0 + 0.15L_t)$. Let $X = \min(T, C)$. If $T < C$, then $\Delta = 1$; otherwise $\Delta = 0$. Finally, let $V = \min(V_1, T, C)$ and let Γ be the indicator of treatment discontinuation before the time to failure or censoring; i.e., if $V = V_1$, then $\Gamma = 1$; otherwise $\Gamma = 0$. The observed data are $(X_i, \Delta_i, V_i, \Gamma_i, \bar{H}_{X_i,i})$ for $i = 1, \ldots, n$. We consider $\psi^* \in \{-0.5, 0, 0.5\}$. From our data-generating mechanism, 50–58% of observations are censored, and 70–80% of treatment discontinuation times are observed before the time to failure or censoring.

We consider the following estimators of ψ^*: (i) a naive estimator $\hat{\psi}_{\mathrm{naive}}$ obtained by solving (4) with T in $U(\psi) = \int_0^T \exp(\psi^* A_u)\,\mathrm{d}u$ replaced by X; (ii) an inverse probability of weighting estimator $\hat{\psi}_{\mathrm{msm}}$ for the Cox marginal structural model in continuous time (Yang et al., 2018); (iii) a simple inverse probability of censoring weighting estimator $\hat{\psi}_{\mathrm{ipcw}}$ obtained by solving $P_n[\{\hat{K}_C(T \mid \bar{H}_T)\}^{-1}\Delta \int c(\bar{H}_u)U(\psi)\,\mathrm{d}M_V(u)] = 0$; and (iv) the proposed doubly robust estimator $\hat{\psi}_{\mathrm{dr}}$ obtained by solving (7) with $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ reduced to a tractable function $E\{U(\psi) \mid$

Table 1. *Simulation results: bias, standard deviation, root mean squared error, and coverage rate of* 95% *confidence intervals for* $\exp(\psi^*)$ *over* 1000 *simulated datasets*

			$\psi^* = -0.5$			$\psi^* = 0$			$\psi^* = 0.5$		
			Bias	SE	CR	Bias	SE	CR	Bias	SE	CR
Scenario 1	$\hat{\psi}_{\text{naive}}$	c	0.06	0.048	76.8	0.02	0.069	95.6	−0.06	0.112	92.4
		c^{opt}	0.05	0.043	78.4	0.02	0.063	95.0	−0.05	0.107	91.8
	$\hat{\psi}_{\text{ipcw}}$	c	−0.01	0.089	95.2	−0.02	0.123	97.2	−0.02	0.191	95.6
		c^{opt}	−0.01	0.070	96.4	−0.02	0.095	97.0	−0.02	0.148	95.6
	$\hat{\psi}_{\text{dr}}$	c	0.00	0.053	95.2	−0.00	0.076	96.8	−0.01	0.125	95.4
		c^{opt}	0.00	0.049	95.4	−0.00	0.071	96.0	−0.00	0.118	94.8
	$\hat{\psi}_{\text{msm}}$	—	−0.00	0.050	95.8	0.00	0.081	96.4	0.00	0.148	95.2
	$\hat{\psi}_{\text{disc}}$	—	−0.37	0.041	0.0	−0.61	0.055	0.0	−1.01	0.092	0.6
Scenario 2	$\hat{\psi}_{\text{naive}}$	c	0.22	0.065	4.8	0.24	0.097	30.4	0.26	0.164	66.0
		c^{opt}	0.22	0.066	5.4	0.24	0.097	31.8	0.26	0.163	68.2
	$\hat{\psi}_{\text{ipcw}}$	c	0.16	0.098	62.4	0.23	0.140	64.4	0.33	0.239	79.6
		c^{opt}	0.16	0.098	62.2	0.23	0.140	65.8	0.33	0.234	79.4
	$\hat{\psi}_{\text{dr}}$	c	0.01	0.048	95.0	0.00	0.070	96.4	0.00	0.115	95.4
		c^{opt}	0.01	0.048	95.4	0.00	0.070	96.6	0.00	0.115	95.2
	$\hat{\psi}_{\text{msm}}$	—	0.13	0.069	54.4	−0.40	0.051	57.6	0.36	0.217	75.6
	$\hat{\psi}_{\text{disc}}$	—	−0.25	0.035	0.0	0.22	0.118	0.0	−0.72	0.092	1.0

Scenario 1, the treatment discontinuation model is correctly specified; Scenario 2, the treatment discontinuation model is misspecified; SE, standard error; CR, coverage rate of 95% confidence intervals.

$\bar{H}_0\}$. Note that $\hat{\psi}_{\text{ipcw}}$ is the special case of $\hat{\psi}_{\text{dr}}$ with $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ misspecified as zero. Moreover, to demonstrate the effect of data discretization, we include the discrete-time g-estimator $\hat{\psi}_{\text{disc}}$ applied to the pre-processed data with grid size 51. The details for $\hat{\psi}_{\text{msm}}$ and $\hat{\psi}_{\text{disc}}$ are presented in the Supplementary Material. For estimators requiring a choice of $c(\bar{H}_u)$, we compare the simple choice $c(\bar{H}_u) = A_{u-}$ and the optimal choice $c^{\text{opt}}(\bar{H}_u)$ in (5), where $E\{\partial \dot{U}_u(\psi)/\partial\psi \mid \bar{H}_u, V = u\} = E(V - u \mid \bar{H}_u, V \geqslant u)$. We approximate $E(V - u \mid \bar{H}_u, V \geqslant u)$ by the mean of the exponential distribution with rate $\hat{\lambda}_V(u)$ and assume that $\text{var}\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ is a constant, which is common practice in the generalized estimating equation literature. We approximate $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ by regressing $\hat{K}_C(T \mid \bar{H}_T)^{-1}\Delta U(\psi)$ on (X_0, L_0). To evaluate the double robustness, we consider two specifications for the hazard of treatment discontinuation: (a) the true proportional hazards model, and (b) a misspecified Kaplan–Meier model (Kaplan & Meier, 1958). In calculating the censoring weights, we specify the censoring model as the true proportional hazards model. We assess the impact of misspecification of the censoring model in the Supplementary Material. For standard errors, we consider the delete-a-group jackknife variance estimator with 500 groups (Kott, 1998).

Table 1 summarizes the simulation results with $n = 1000$. The naive estimator $\hat{\psi}_{\text{naive}}$ is biased, and its bias becomes larger as $|\psi^*|$ increases. In scenario 1, where the treatment process model is correctly specified, $\hat{\psi}_{\text{ipcw}}$, $\hat{\psi}_{\text{dr}}$ and $\hat{\psi}_{\text{msm}}$ show small biases across all scenarios with different values of ψ^*. Note that $\hat{\psi}_{\text{ipcw}}$ is a special case of the proposed estimator with $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ misspecified as zero. This demonstrates that the proposed estimator is robust to misspecification of $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ given that the treatment process model is correctly specified. If additionally $E\{U(\psi) \mid \bar{H}_u, V \geqslant u\}$ is well approximated, $\hat{\psi}_{\text{dr}}$ achieves gains in estimation efficiency over $\hat{\psi}_{\text{ipcw}}$. Moreover, $\hat{\psi}_{\text{dr}}$ with c^{opt} is more efficient than with c. In scenario 1, $\hat{\psi}_{\text{dr}}$ has smaller standard errors than $\hat{\psi}_{\text{msm}}$. This is because $\hat{\psi}_{\text{msm}}$ involves weighting

Table 2. *Results of the effect of oral anticoagulant therapy on the composite outcome;* $\exp(\psi^*)$ *is the causal estimand*

	Est	SE	CI	p-value
Naive method	0.68	0.176	(0.34, 1.03)	0.07
Proposed method	0.64	0.179	(0.29, 0.99)	0.04

Est, estimate of $\exp(\psi^*)$; SE, standard error; CI, 95% confidence interval.

directly by the inverse of the propensity score, whereas $\hat{\psi}_{\mathrm{dr}}$ utilizes the propensity score not in the form of inverse weights and therefore avoids the possibly large variability due to weighting. In scenario 2, where the treatment process model is misspecified, $\hat{\psi}_{\mathrm{ipcw}}$ and $\hat{\psi}_{\mathrm{msm}}$ show large biases; however, $\hat{\psi}_{\mathrm{dr}}$ still has small biases, confirming its double robustness. The jackknife variance estimation performs well for $\hat{\psi}_{\mathrm{dr}}$ and produces coverage rates close to the nominal level. Large biases in the discrete-time g-estimator $\hat{\psi}_{\mathrm{disc}}$ illustrate the consequences of data pre-processing for the subsequent analysis.

6. Application to the GARFIELD data

We analyse data from the Global Anticoagulant Registry in the FIELD with Atrial Fibrillation, GARFIELD-AF, registry study, an observational study of patients newly diagnosed with atrial fibrillation; see the study website at `http://www.garfieldregistry.org/` for details. Our analysis includes 22 811 patients who were enrolled between April 2013 and August 2016, and received oral anticoagulant therapy for stroke prevention. The goal is to investigate the effect of discontinuation of oral anticoagulant therapy in patients with atrial fibrillation. The primary endpoint is the composite clinical outcome, including death, non-haemorrhagic stroke, systemic embolism and myocardial infarction. Treatment discontinuation at time t is defined as treatment being stopped at time t and never restarted afterwards. In our study, 9.5% of patients discontinued oral anticoagulant therapy over a median follow-up of 710 days with an interquartile range of (487, 731) days; 43.8% of discontinuations were within the first four months of beginning treatment. Among patients who discontinued treatment, 512 stopped the treatment for more than seven days and then went back on treatment. We censor these patients at the time of restarting treatment. This censoring mechanism is not likely to be completely at random, because patients with poor prognosis may be more likely to restart. We assume a dependent censoring mechanism and use inverse probability of censoring weighting.

To answer the clinical question of interest, we consider the structural failure time model $U(\psi^*) = \int_0^T \exp(\psi^* A_u)\,\mathrm{d}u$. Under this model, if a patient had been on treatment continuously, $T = U(\psi^*)\exp(-\psi^*)$, so $U(\psi^*)\{\exp(-\psi^*) - 1\}$ is the time gained or lost while on treatment. We focus on estimating the multiplicative factor $\exp(\psi^*)$. Table 2 reports the results obtained from using the naive estimator and the proposed doubly robust estimator as described in § 5. The details of the nuisance models are given in the Supplementary Material. Although the effect sizes may be a little different between the naive analysis and the proposed analysis, qualitatively they all suggest that treatment is beneficial for prolonging the time to clinical events, and thus that treatment discontinuation is harmful. If a patient had been on treatment continuously, the time to clinical outcomes would have been $\exp(-\hat{\psi}) = 1/0.64 = 1.56$ times longer than if the patient had never received treatment. Importantly, the proposed analysis is designed to address the well-formulated question for investigating the effect of treatment discontinuation.

7. DISCUSSION

The proposed framework of structural failure time models can be used to adjust for time-varying confounding and selection bias with irregularly spaced observations under the three assumptions of no unmeasured confounders, ignorability of censoring, and positivity. As discussed previously, the first and second assumptions hold in the scenario of adjusting for all variables that are related to both treatment discontinuation and outcome, and all variables that are related to both censoring and outcome. Although essential, these assumptions are not verifiable based on the observed data, but rely on subject-matter experts' assessments of their plausibility. Future work will investigate the sensitivity to these assumptions using the methods of Yang & Lok (2017). The third assumption is that all subjects have nonzero probabilities of staying on study before the failure time; it requires the absence of predictors that are deterministic in relation to censoring and outcome. Practitioners should carefully examine the question at hand to eliminate deterministic violations of positivity.

Our framework can also be extended in the following directions. First, the proposed doubly robust estimator still relies on a correct specification of the censoring mechanism. If the censoring model is misspecified, the proposed estimator may be biased; see the additional simulation results in the Supplementary Material. It would be interesting to construct an improved estimator that is multiply robust in the sense that it is consistent in the union of the three models (Molina et al., 2017). Second, it is critical to derive test procedures for evaluating the goodness-of-fit of the treatment effect model (Yang & Lok, 2016).

ACKNOWLEDGEMENT

We have benefited from the comments from two reviewers and Anastasio A. Tsiatis. The first author was partially supported by the U.S. National Science Foundation and National Cancer Institute.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theoretical results and additional simulations. The `R` package implementing the methods in this article is available at `https://github.com/shuyang1987/contTimeCausal`.

REFERENCES

ANDERSEN, P. K., BORGAN, O., GILL, R. D. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

ATKINSON, K. E. (1989). *An Introduction to Numerical Analysis*. New York: Wiley.

BANG, H. & ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–73.

BICKEL, P. J., KLAASSEN, C., RITOV, Y. & WELLNER, J. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore, Maryland: Johns Hopkins University Press.

CAO, W., TSIATIS, A. A. & DAVIDIAN, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–34.

COX, D. R. & OAKES, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

DANIEL, R., COUSENS, S., DE STAVOLA, B., KENWARD, M. & STERNE, J. (2013). Methods for dealing with time-dependent confounding. *Statist. Med.* **32**, 1584–618.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

EFRON, B. & STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–96.

HERNÁN, M. A., BRUMBACK, B. & ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Am. Statist. Assoc.* **96**, 440–8.

Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M. & Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol. Drug Safety* **14**, 477–91.
Joffe, M. M. (2001). Administrative and artificial censoring in censored regression models. *Statist. Med.* **20**, 2287–304.
Joffe, M. M., Yang, W. P. & Feldman, H. (2012). G-estimation and artificial censoring: Problems, challenges, and applications. *Biometrics* **68**, 275–86.
Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.
Kott, P. S. (1998). Using the delete-a-group jackknife variance estimator in practice. In *Proc. Surv. Res. Meth. Sect., ASA*. Alexandria, Virginia: American Statistical Association, pp. 763–8.
Lok, J., Gill, R., Van Der Vaart, A. & Robins, J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statist. Neer.* **58**, 271–95.
Lok, J. J. (2008). Statistical modeling of causal effects in continuous time. *Ann. Statist.* **36**, 1464–507.
Lok, J. J. (2017). Mimicking counterfactual outcomes to estimate causal effects. *Ann. Statist.* **45**, 461–99.
Lok, J. J. & DeGruttola, V. (2012). Impact of time to start treatment following infection with application to initiating HAART in HIV-positive patients. *Biometrics* **68**, 745–54.
Lunceford, J. K. & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statist. Med.* **23**, 2937–60.
Mark, S. D. & Robins, J. M. (1993a). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statist. Med.* **12**, 1605–28.
Mark, S. D. & Robins, J. M. (1993b). A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial. *Contr. Clin. Trials* **14**, 79–97.
Molina, J., Rotnitzky, A., Sued, M. & Robins, J. (2017). Multiple robustness in factorized likelihood models. *Biometrika* **104**, 561–81.
R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79**, 321–34.
Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statist. Sci.* **22**, 544–59.
Robins, J. M. (1998a). Correction for non-compliance in equivalence trials. *Statist. Med.* **17**, 269–302.
Robins, J. M. (1998b). Structural nested failure time models. In *The Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds. Chichester: Wiley, pp. 4372–89.
Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer, pp. 95–133.
Robins, J. M. (2002). Analytic methods for estimating HIV-treatment and cofactor effects. In *Methodological Issues in AIDS Behavioral Research*. New York: Springer, pp. 213–88.
Robins, J. M., Blevins, D., Ritter, G. & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319–36.
Robins, J. M. & Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *J. Am. Statist. Assoc.* **89**, 737–49.
Robins, J. M., Hernan, M. A. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–60.
Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
Robins, J. M. & Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun. Statist.* A **20**, 2609–31.
Rotnitzky, A., Bergesio, A. & Farall, A. (2009). Analysis of quality-of-life adjusted failure time data in the presence of competing, possibly informative, censoring mechanisms. *Lifetime Data Anal.* **15**, 1–23.
Rotnitzky, A. & Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–20.
Rotnitzky, A. & Vansteelandt, S. (2015). Double-robust methods. In *Handbook of Missing Data Methodology*, A. Tsiatis & G. Verbeke, eds. Boca Raton, Florida: CRC Press, pp. 185–212.
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.
Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Statist. Assoc.* **94**, 1096–120.
Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
Van Der Laan, M. J., Hubbard, A. E. & Robins, J. M. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *J. Am. Statist. Assoc.* **97**, 494–507.
van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Emprical Processes: With Applications to Statistics*. New York: Springer.

WITTEMAN, J. C., D'AGOSTINO, R. B., STIJNEN, T., KANNEL, W. B., COBB, J. C., DE RIDDER, M. A., HOFMAN, A. & ROBINS, J. M. (1998). G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *Am. J. Epidemiol.* **148**, 390–401.

YANG, S. & LOK, J. J. (2016). A goodness-of-fit test for structural nested mean models. *Biometrika* **103**, 734–41.

YANG, S. & LOK, J. J. (2017). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statist. Sinica* **28**, 1703–23.

YANG, S., TSIATIS, A. A. & BLAZING, M. (2018). Modeling survival distribution as a function of time to treatment discontinuation: A dynamic treatment regime approach. *Biometrics* **74**, 900–9.

YOUNG, J. G., HERNÁN, M. A., PICCIOTTO, S. & ROBINS, J. M. (2010). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Anal.* **16**, 71–84.

ZHANG, M., JOFFE, M. M. & SMALL, D. S. (2011). Causal inference for continuous-time processes when covariates are observed only at discrete times. *Ann. Statist.* **39**, 131–73.

[*Received on* 20 *August* 2018. *Editorial decision on* 9 *April* 2019]

Biometrika (2018), **xx**, x, *pp*. 1–16
Printed in Great Britain

Supplementary material for "Semiparametric estimation of structural failure time model in continuous-time processes"

BY S. YANG
Department of Statistics, North Carolina State University, North Carolina 27695, U.S.A.
syang24@ncsu.edu

K. PIEPER
Duke Clinical Research Institute, Duke University, North Carolina 27705, U.S.A.
karen.pieper@duke.edu

AND F. COOLS
Department of Cardiology, AZ Klina, Augustijnslei 100, 2930 Brasschaat, Belgium.
frank.cools@klina.be

SUPPLEMENTARY MATERIAL

S1. A LEMMA

We provide a lemma for the martingale process, which is useful in our derivation later.

Consider the Hilbert space $\mathcal{H}$ of all p-dimensional, mean-zero finite variance measurable functions of F, $h(F)$, equipped with the covariance inner product $< h_1, h_2 >= E\left\{h_1(F)^{\mathrm{T}} h_2(F)\right\}$ and the norm $||h|| = \left[E\left\{h(F)^{\mathrm{T}} h(F)\right\}\right]^{1/2} < \infty$.

LEMMA S1. *Under Assumption 1, $M_V(t)$ is a martingale with respect to the filtration $\sigma\{\overline{H}_t, U(\psi^*)\}$. By Proposition II.4.1 in Andersen et al. (1993), $M_V(t)$ has a unique compensator $< M_V(t) >= \int_0^t \lambda_V(u \mid \overline{H}_u) Y_V(u)\mathrm{d}u$. If $g_1(\cdot)$ and $g_2(\cdot)$ are bounded $\sigma\{\overline{H}_t, U(\psi^*)\}$-predictable processes, then $< \int_0^t g_1(u)\mathrm{d}M_V(u), \int_0^t g_2(u)\mathrm{d}M_V(u) >$ exists, and*

$$< \int_0^t g_1(u)\mathrm{d}M_V(u), \int_0^t g_2(u)\mathrm{d}M_V(u) >= \int_0^t g_1(u)g_2(u)\lambda_V(u \mid \overline{H}_u) Y_V(u)\mathrm{d}u. \qquad \text{(S1)}$$

S2. PROOF OF $\lambda_V\{t \mid \overline{H}_t, U(\psi^*)\} = \lambda_V(t \mid \overline{H}_t)$

It suffices to show that $\lambda_V\{t \mid \overline{H}_t, U(\psi^*)\} = \lambda_V(t \mid \overline{H}_t, U)$. We obtain

$$\begin{aligned}
&\lambda_V\left(t \mid \overline{H}_t, U\right) = \lim_{h\to 0} h^{-1} P\left(t \leqslant V < t+h, \Gamma = 1 \mid V \geqslant t, \overline{H}_t, U\right) \\
&= \lim_{h\to 0} h^{-1} \frac{P\left(U \mid t \leqslant V < t+h, \Gamma = 1, \overline{H}_t\right) P(t \leqslant V < t+h, \Gamma = 1 \mid V \geqslant t, \overline{H}_t)}{P\left(U \mid V \geqslant t, \Gamma = 1, \overline{H}_t\right)} \\
&= \lim_{h\to 0} h^{-1} \frac{P\left\{U(\psi^*) \mid t \leqslant V < t+h, \Gamma = 1, \overline{H}_t\right\} P(t \leqslant V < t+h, \Gamma = 1 \mid V \geqslant t, \overline{H}_t)}{P\left\{U(\psi^*) \mid V \geqslant t, \Gamma = 1, \overline{H}_t\right\}} \\
&= \lim_{h\to 0} h^{-1} P\left\{t \leqslant V < t+h, \Gamma = 1 \mid V \geqslant t, \overline{H}_t, U(\psi^*)\right\} \\
&= \lambda_V\{t \mid \overline{H}_t, U(\psi^*)\},
\end{aligned}$$

where the second equality follows by the Bayes rule, and the third equality follows by Model (1) which entails that the distributions of $(U, \overline{H}_t)$ and $\{U(\psi^*), \overline{H}_t\}$ are the same.

S3. IDENTIFICATION OF $\psi \in \mathcal{R}^p$ UNDER ASSUMPTION 1

Under Assumption 1, $M_V(t) = N_V(t) - \int_0^t \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u$ is a martingale with respect to the filtration $\sigma\{\overline{H}_t, U(\psi^*)\}$. Then, for any $c(\overline{H}_t) \in \mathcal{R}^p$ and $t > 0$,

$$E\left\{c(\overline{H}_t) U(\psi^*) \mathrm{d}M_V(t)\right\} = 0. \tag{S2}$$

Suppose that (S2) holds for ψ^{*1} and ψ^{*2}; i.e., for any $c(\overline{H}_t) \in \mathcal{R}^p$ and $t > 0$, $E\left[c(\overline{H}_t)\left\{U(\psi^{*1}) - U(\psi^{*2})\right\} \mathrm{d}M_V(t)\right] = 0$. To reflect the dependence of $U(\psi^{*1}) - U(\psi^{*2})$ on $\left(\overline{A}_T, \overline{L}_T\right)$, denote $\varphi\left(\overline{A}_T, \overline{L}_T\right) = U(\psi^{*1}) - U(\psi^{*2}) = \int_0^T \exp[\{\psi_1^{*1} + \psi_2^{*1\mathrm{T}} g(L_u)\} A_u] \mathrm{d}u - \int_0^T \exp[\{\psi_1^{*2} + \psi_2^{*2\mathrm{T}} g(L_u)\} A_u] \mathrm{d}u$. Then, for any $c(\overline{H}_t) \in \mathcal{R}^p$ and $t > 0$, we have $E\left\{c(\overline{H}_t) \varphi\left(\overline{A}_T, \overline{L}_T\right) \mathrm{d}M_V(t)\right\} = 0$. This implies that $\varphi\left(\overline{A}_T, \overline{L}_T\right)$ is independent of $M_V(t)$ conditional on $(\overline{H}_t, V > t)$ for all $\overline{H}_t$ and $t > 0$. Therefore, $\varphi\left(\overline{A}_T, \overline{L}_T\right)$ must not depend on $\overline{A}_T$, and therefore ψ^{*1} must equal ψ^{*2}. Consequently, ψ^* is uniquely identified from (S2).

S4. PROOF OF THEOREM 1

To motivate the concept of the nuisance tangent space for a semiparametric model, we first consider a parametric model $f(F; \psi, \theta)$, where ψ is a p-dimensional parameter of interest, and θ is an q-dimensional nuisance parameter. The score vectors of ψ and θ are $S_\psi(F) = \partial \log f(F; \psi, \theta^*)/\partial\psi$ and $S_\theta(F) = \partial \log f(F; \psi^*, \theta)/\partial\theta$, respectively, both evaluated at the true value (ψ^*, θ^*). For this parametric model, the nuisance tangent space Λ is the linear space in $\mathcal{H}$ spanned by the nuisance score vector $S_\theta(F)$. In a semiparametric model, the nuisance parameter θ may be infinite-dimensional. The nuisance tangent space Λ is defined as the mean squared closure of the nuisance tangent spaces under any parametric submodel. An important fact is that the orthogonal complement of the nuisance tangent space $\Lambda^\perp$ contains the influence functions for regular asymptotically linear estimators of ψ.

First, we characterize the semiparametric likelihood function based on a single observable F. Because the transformation of F to $\{U(\psi^*), \overline{H}_T\}$ is one-to-one, the likelihood function based on F becomes

$$f_F\left(T, \overline{H}_T\right) = \left\{\frac{\partial U(\psi^*)}{\partial T}\right\} f_{\{U(\psi^*), \overline{H}_T\}}\{U(\psi^*), \overline{H}_T\}, \tag{S3}$$

where $\partial U(\psi^*)/\partial T = \exp\left[A_T\{\psi_1^* + \psi_2^{*\mathrm{T}} g(L_T)\}\right]$. Let $0 = v_0 < v_1 < \cdots < v_M$ be the observed times to treatment discontinuation among the n subjects. We further express (S3) as

$$\begin{aligned} f_F\left(T, \overline{H}_T; \psi^*, \theta\right) = \left\{\frac{\partial U(\psi^*)}{\partial T}\right\} & f\left\{U(\psi^*); \theta_1\right\} \prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{H}_{v_k - 1}, U(\psi^*), T > v_k; \theta_2\right\} \\ & \times \prod_{v=v_1}^{v_M} f\left\{A_{v_k} \mid \overline{H}_{v_k - 1}, U(\psi^*), T > v_k; \theta_3\right\} \end{aligned}$$

$$= \left\{ \frac{\partial U(\psi^*)}{\partial T} \right\} f\left\{U(\psi^*); \theta_1\right\} \prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{H}_{v_k-1}, U(\psi^*), T > v_k; \theta_2\right\}$$
$$\times \prod_{v=v_1}^{v_M} f\left(A_{v_k} \mid \overline{H}_{v_k-1}, T > v_k; \theta_3\right), \tag{S4}$$

where the second equality follows from Assumption 1, $f\left\{U(\psi^*)\right\}$, $f\left\{L_{v_k} \mid \overline{H}_{v_k-1}, U(\psi^*), T > v_k\right\}$, and $f\left(A_{v_k} \mid \overline{H}_{v_k-1}, T > v_k\right)$ are completely unspecified, and $\theta = (\theta_1, \theta_2, \theta_3)$ is a vector of infinite-dimensional nuisance parameters.

Let Λ_k be the nuisance tangent space for θ_k, for $k = 1, 2, 3$. We now characterize Λ_k.

For the nuisance parameter θ_1, $f\left\{U(\psi^*); \theta_1\right\}$ is a nonparametric model indexed by θ_1, i.e., $f\left\{U(\psi^*); \theta_1\right\}$ is a non-negative function and satisfies $\int f\left(v; \theta_1\right) \mathrm{d}v = 1$. Following § 4.4 of Tsiatis (2006), the tangent space regarding θ_1 is the set of all vector $s\left\{U(\psi^*)\right\} \in \mathcal{R}^p$ with $E\left[s\left\{U(\psi^*)\right\}\right] = 0$. Thus, the tangent space of θ_1 is

$$\Lambda_1 = \left\{s\left\{U(\psi^*)\right\} \in \mathcal{R}^p : E\left[s\left\{U(\psi^*)\right\}\right] = 0\right\}.$$

For the nuisance parameter θ_2, $\prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{H}_{v_k-1}, U(\psi^*), T > v_k; \theta_2\right\}$ is a nonparametric model indexed by θ_2. To obtain the nuisance tangent space of θ_2, following the same derivation as for θ_1, the score function of θ_2 is of the form $\sum_{k=1}^{M} S\left\{L_{v_k}, \overline{H}_{v_k-1}, U(\psi^*)\right\}$, where $E[S\left\{L_{v_k}, \overline{H}_{v_k-1}, U(\psi^*)\right\} \mid \overline{H}_{v_k-1}, U(\psi^*), T > v_k] = 0$. Thus, the tangent space of θ_2 is

$$\Lambda_2 = \sum_{k=1}^{M} \Big\{ S\left\{L_{v_k}, \overline{H}_{v_k-1}, U(\psi^*)\right\} \in \mathcal{R}^p :$$
$$E\left[S\left\{L_{v_k}, \overline{H}_{v_k-1}, U(\psi^*)\right\} \mid \overline{H}_{v_k-1}, U(\psi^*), T > v_k\right] = 0 \Big\}.$$

For the nuisance parameter θ_3, $\prod_{k=1}^{M} f\left(A_{v_k} \mid \overline{L}_{v_k-1}, \overline{A}_{v_k-1}, T > v_k; \theta_3\right)$ can be equivalently expressed as the likelihood based on the data $(V, \Gamma, \overline{H}_V)$ and the hazard function $\lambda_V(t \mid \overline{H}_t)$:

$$f_{(V,\Gamma,\overline{H}_V)}(V, \Gamma, \overline{H}_V) = \lambda_V(V \mid \overline{H}_V)^{\Gamma} \exp\left\{ - \int_0^V \lambda_V(u \mid \overline{H}_u) \mathrm{d}u \right\}$$
$$\times \left\{ f_{T|\overline{H}_T}(V \mid \overline{H}_V) \right\}^{1-\Gamma} \left\{ \int_V^{\infty} f_{T|\overline{H}_T}(u \mid \overline{H}_u) \mathrm{d}u \right\}^{\Gamma}.$$

Following Tsiatis (2006), the tangent space of θ_3 is

$$\Lambda_3 = \left\{ \int h_u(\overline{H}_u) \mathrm{d}M_V(u) : \; h_u(\overline{H}_u) \in \mathcal{R}^p \right\}.$$

Moreover, it is easy to show that Λ_1, Λ_2 and Λ_3 are mutually orthogonal subspaces. Then, $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$, where $\oplus$ denotes a direct sum.

Now, let

$$\Lambda_3^* = \left\{ \int h_u\{U(\psi^*), \overline{H}_u\} \mathrm{d}M_V(u) : \; h_u\{U(\psi^*), \overline{H}_u\} \in \mathcal{R}^p \right\}.$$

Because the tangent space $\Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3^*$ is that for a nonparametric model; i.e., a model that allows for all densities of F, and because the tangent space for a nonparametric model is the entire Hilbert space, we obtain $\mathcal{H} = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3^*$. Because $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$, this implies

that $\Lambda_3 \subset \Lambda_3^*$. Also, the orthogonal complement $\Lambda^\perp$ must be orthogonal to $\Lambda_1 \oplus \Lambda_2$, so $\Lambda^\perp$ must belong to Λ_3^* and be orthogonal to Λ_3. This means that $\Lambda^\perp$ consists of all elements of Λ_3^* that are orthogonal to Λ_3.

To characterize $\Lambda^\perp$, for any $\int h_u\{U(\psi^*), \overline{H}_u\}\mathrm{d}M_V(u) \in \Lambda_3^*$, we obtain its projection onto $\Lambda_3^\perp$. To find the projection, we derive $h_u^*(\overline{H}_u)$ so that

$$\left[\int h_u\{U(\psi^*), \overline{H}_u\}\mathrm{d}M_V(u) - \int h_u^*(\overline{H}_u)\mathrm{d}M_V(u)\right] \in \Lambda_3^\perp.$$

Therefore, we have

$$E\left(\int \left[h_u\{U(\psi^*), \overline{H}_u\} - h_u^*(\overline{H}_u)\right] \mathrm{d}M_V(u) \times \int h_u(\overline{H}_u)\mathrm{d}M_V(u)\right) = 0, \tag{S5}$$

for any $h_u(\overline{H}_u)$. By Lemma S1, (S5) becomes

$$\begin{aligned} &E\left(\int \left[h_u\{U(\psi^*), \overline{H}_u\} - h_u^*(\overline{H}_u)\right] h_u(\overline{H}_u)\lambda_V(u \mid \overline{H}_u)Y_V(u)\mathrm{d}u\right) \\ &= E\left(\int E\left(\left[h_u\{U(\psi^*), \overline{H}_u\} - h_u^*(\overline{H}_u)\right] Y_V(u) \mid \overline{H}_u\right) h_u(\overline{H}_u)\lambda_V(u \mid \overline{H}_u)\mathrm{d}u\right) = 0 \end{aligned}$$

for any $h_u(\overline{H}_u)$. Because $h_u(\overline{H}_u)$ is arbitrary, we must have

$$E\left(\left[h_u\{U(\psi^*), \overline{H}_u\} - h_u^*(\overline{H}_u)\right] Y_V(u) \mid \overline{H}_u\right) = 0. \tag{S6}$$

Solving (S6) for $h_u^*(\overline{H}_u)$, we obtain

$$E\left[h_u\{U(\psi^*), \overline{H}_u\}Y_V(u) \mid \overline{H}_u\right] = h_u^*(\overline{H}_u)E\left\{Y_V(u) \mid \overline{H}_u\right\},$$

or

$$h_u^*(\overline{H}_u) = \frac{E\left[h_u\{U(\psi^*), \overline{H}_u\}Y_V(u) \mid \overline{H}_u\right]}{E\left\{Y_V(u) \mid \overline{H}_u\right\}} = E\left[h_u\{U(\psi^*), \overline{H}_u\} \mid \overline{H}_u, V \geqslant u\right].$$

Therefore, the space orthogonal to the nuisance tangent space is given by

$$\Lambda^\perp = \left\{\int \left(h_u\{U(\psi^*), \overline{H}_u\} - E\left[h_u\{U(\psi^*), \overline{H}_u\} \mid \overline{H}_u, V \geqslant u\right]\right) \mathrm{d}M_V(u) : h_u\{U(\psi^*), \overline{H}_u\} \in \mathcal{R}^p\right\}.$$

S5. THE OPTIMAL FORM $c^{\mathrm{opt}}(\overline{H}_u)$

We obtain the optimal form of $c(\overline{H}_u)$ by projecting the score function $S_\psi(F)$ onto

$$\Lambda_0^\perp = \left\{G(\psi^*; F, c) = \int_0^\infty c(\overline{H}_u)\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u) : c(\overline{H}_u) \in \mathcal{R}^p\right\}.$$

We first characterize the projection of any $B(F) \in \mathcal{H}$ onto $\Lambda_0^\perp$. For ease of notation, we may suppress the dependence of F of random variables if there is no ambiguity.

THEOREM S1 (PROJECTION). *For any $B = B(F) \in \mathcal{H}$, the projection of B onto $\Lambda_0^{\perp}$ is*

$$\prod\left(B \mid \Lambda_0^{\perp}\right) = \int \left[E\left\{B\dot{U}_u(\psi^*) \mid \overline{H}_u, V = u\right\} - E\left\{B\dot{U}_u(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]$$
$$\times \left[\mathrm{var}\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]^{-1} \left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u), \quad \text{(S7)}$$

where $\dot{U}_u(\psi) = U(\psi) - E\{U(\psi) \mid \overline{H}_u, V \geqslant u\}$.

Proof. Let $G(F)$ be the quantity in the right hand side of (S7). To show that $\prod\left(B \mid \Lambda_0^{\perp}\right) = G(F)$, we must show that $B - G \in \Lambda_0$. Toward that end, we show that for any $\widetilde{G}(F) \in \Lambda_0^{\perp}$, $(B - G) \perp\!\!\!\perp \widetilde{G}$. Specifically, we need to show that for any $\widetilde{G}(F) = \int_0^{\infty} \widetilde{c}(\overline{H}_u)\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u)$, $E\left\{(B - G)\widetilde{G}\right\} = 0$. We now verify that $E\left(B\widetilde{G}\right) = E\left(G\widetilde{G}\right)$ by the following calculation.

Firstly, we obtain

$$E\left(G\widetilde{G}\right) = E\left(< G, \widetilde{G} >\right)$$
$$= E\int \widetilde{c}(\overline{H}_u)\left[E\left\{BU(\psi^*) \mid \overline{H}_u, V = u\right\} - E\left\{BU(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \quad \text{(S8)}$$
$$\times \left[\mathrm{var}\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]^{-1} \left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]^2 \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u$$
$$= E\int \widetilde{c}(\overline{H}_u)\left[E\left\{BU(\psi^*) \mid \overline{H}_u, V = u\right\} - E\left\{BU(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u.$$

Secondly, we obtain

$$E\left(B\widetilde{G}\right) = E\int \widetilde{c}(\overline{H}_u) B\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, T \geqslant u\right\}\right] \mathrm{d}M_V(u)$$
$$= E\int \widetilde{c}(\overline{H}_u) B\dot{U}_u(\psi^*) \mathrm{d}N_V(u) - E\int_0^{\infty} \widetilde{c}(\bar{V}_u) B\dot{U}_u(\psi^*) \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u \quad \text{(S9)}$$
$$= E\int \widetilde{c}(\overline{H}_u)\left[E\left\{B\dot{U}_u(\psi^*) \mid \overline{H}_u, V = u\right\} - E\left\{B\dot{U}_u(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u,$$

where the last equality follows because

$$E\int \widetilde{c}(\overline{H}_u) B\dot{U}_u(\psi^*) \mathrm{d}N_V(u) = E\int \widetilde{c}(\overline{H}_u) E\left\{B\dot{U}_u(\psi^*) \mathrm{d}N_V(u) \mid \overline{H}_u, V \geqslant u\right\}$$
$$= E\int \widetilde{c}(\overline{H}_u) E\left\{B\dot{U}_u(\psi^*) I\left(u \leqslant V \leqslant u + \mathrm{d}u, \Gamma = 1\right) \mid \overline{H}_u, V \geqslant u\right\}$$
$$= E\int \widetilde{c}(\overline{H}_u) E\left\{B\dot{U}_u(\psi^*) \mid \overline{H}_u, V = u\right\} \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u,$$

and

$$E\int \widetilde{c}(\overline{H}_u) B\dot{U}_u(\psi^*) \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u$$
$$= E\int \widetilde{c}(\bar{V}_u) E\left\{B\dot{U}_u(\psi^*) \mid \overline{H}_u, V \geqslant u\right\} \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u.$$

Therefore, by (S8) and (S9), $E\left(B\widetilde{G}\right) = E\left(G\widetilde{G}\right)$ for any $\widetilde{G} \in \Lambda_0^{\perp}$, proving (S7).

THEOREM S2. *The optimal form of $c(\overline{H}_u)$ is (5) in the sense that with this form the solution to (4) gives the most precise estimator of ψ^* among all the solutions to (4).*

Proof. We write $G(\psi^*; F, c)$ to emphasize its dependence on $c(\overline{H}_u)$. We derive the optimal form of $c(\overline{H}_u)$ by deriving the most efficient $G(\psi^*; F, c)$ in $\Lambda_0^\perp$, which is $G(\psi^*; F, c^{\text{opt}}) = \prod(S_\psi \mid \Lambda_0^\perp)$.

By Theorem S1, we have

$$G(\psi^*; F, c^{\text{opt}}) = \int \left[E\left\{S_\psi \dot{U}_u(\psi^*) \mid \overline{H}_u, V = u\right\} - E\left\{S_\psi \dot{U}_u(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \times \left[\text{var}\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]^{-1} \left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u). \quad \text{(S10)}$$

Because $E\{\dot{U}_u(\psi) \mid \overline{H}_u, V \geqslant u\} = 0$, taking the derivative of ψ at both sides and using the generalized information equality, we have $E\{S_\psi \dot{U}_u(\psi) \mid \overline{H}_u, V \geqslant u\} + E\{\partial \dot{U}_u(\psi)/\partial\psi \mid \overline{H}_u, V \geqslant u\} = 0$, or equivalently $E\{S_\psi \dot{U}_u(\psi) \mid \overline{H}_u, V \geqslant u\} = -E\{\partial \dot{U}_u(\psi)/\partial\psi \mid \overline{H}_u, V \geqslant u\}$. Similarly, because $E\{\dot{U}_u(\psi) \mid \overline{H}_u, V = u\} = 0$, we have $E\{S_\psi \dot{U}_u(\psi) \mid \overline{H}_u, V = u\} + E\{\partial \dot{U}_u(\psi)/\partial\psi \mid \overline{H}_u, V = u\} = 0$, or equivalently $E\{S_\psi \dot{U}_u(\psi) \mid \overline{H}_u, V = u\} = -E\{\partial \dot{U}_u(\psi)/\partial\psi \mid \overline{H}_u, V = u\}$. Continuing (S10),

$$\begin{aligned} G(\psi^*; F, c^{\text{opt}}) &= -\int_0^\infty \left[E\left\{\partial \dot{U}_u(\psi^*)/\partial\psi \mid \overline{H}_u, V = u\right\} - E\left\{\partial \dot{U}_u(\psi^*)/\partial\psi \mid \overline{H}_u, V \geqslant u\right\}\right] \\ &\quad \times \left[\text{var}\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]^{-1} \left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u) \\ &= -\int_0^\infty E\left\{\partial \dot{U}_u(\psi^*)/\partial\psi \mid \overline{H}_u, V = u\right\} \left[\text{var}\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]^{-1} \\ &\quad \times \left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u). \end{aligned} \quad \text{(S11)}$$

Therefore, by (S11), ignoring the negative sign, $c^{\text{opt}}(\overline{H}_u)$ is given by (5).

S6. PROOF OF THEOREM 2

We show that $E\{G(\psi^*; F, c)\} = 0$ in two cases.

First, if $\lambda_V(t \mid \overline{H}_t)$ is correctly specified, under Assumption 1, $M_V(t)$ is a martingale with respect to the filtration $\sigma\{\overline{H}_t, U(\psi^*)\}$. Because $c(\overline{H}_u)\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right]$ is a $\sigma\{\overline{H}_t, U(\psi^*)\}$-predictable process, $\int_0^t c(\overline{H}_u)\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}\right] \mathrm{d}M_V(u)$ is a martingale for $t \geqslant 0$. Therefore, $E\{G(\psi^*; F, c)\} = 0$.

Second, if $E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\right\}$ is correctly specified but $\lambda_V(t \mid \overline{H}_t)$ is not necessarily correctly specified, let $\lambda_V^*(t \mid \overline{H}_t)$ be the probability limit of the possibly misspecified model. We obtain

$$\begin{aligned} &E\int c(\overline{H}_u)\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi^*\right\}\right]\left\{\mathrm{d}N_V(u) - \lambda_V^*(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u\right\} \\ &= E\int c(\overline{H}_u)\left[U(\psi^*) - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi^*\right\}\right]\left\{\mathrm{d}N_V(u) - \lambda_V(u \mid \overline{H}_u) Y_V(u) \mathrm{d}u\right\} \end{aligned}$$

$$
\begin{aligned}
&+E\int c(\overline{H}_u)\left[U(\psi^*)-E\left\{U(\psi^*)\mid \overline{H}_u, V\geqslant u;\xi^*\right\}\right]\left\{\lambda_V(u\mid \overline{H}_u)-\lambda_V^*(u\mid \overline{H}_u)\right\}Y_V(u)\mathrm{d}u\\
&=0+E\int c(\overline{H}_u)E\left(\left[U(\psi^*)-E\left\{U(\psi^*)\mid \overline{H}_u, V\geqslant u;\xi^*\right\}\right]\mid \overline{H}_u, V\geqslant u\right) \qquad \text{(S12)}\\
&\quad\times\left\{\lambda_V(u\mid \overline{H}_u)-\lambda_V^*(u\mid \overline{H}_u)\right\}Y_V(u)\mathrm{d}u\\
&=0+E\int c(\overline{H}_u)\times 0\times\left\{\lambda_V(u\mid \overline{H}_u)-\lambda_V^*(u\mid \overline{H}_u)\right\}Y_V(u)\mathrm{d}u \qquad \text{(S13)}\\
&=0,
\end{aligned}
$$

where zero in (S12) follows because $\mathrm{d}M_V(u)=\mathrm{d}N_V(u)-\lambda_V(u\mid \overline{H}_u)\mathrm{d}u$ is a martingale with respect to the filtration $\sigma\{\overline{H}_t, U(\psi^*)\}$, and zero in (S13) follows because $E\left\{U(\psi^*)\mid \overline{H}_u, V\geqslant u\right\}$ is correctly specified and therefore, $E\left\{U(\psi^*)\mid \overline{H}_u, V\geqslant u;\xi^*\right\}=E\left\{U(\psi^*)\mid \overline{H}_u, V\geqslant u\right\}$.

S7. Proof that $\widetilde{U}(\psi^*)$ and $\Delta(\psi^*)$ are computable

If $T\leqslant C$, because $U(\psi^*)$ and $C(\psi^*)$ are observable, $\widetilde{U}(\psi^*)$ and $\Delta(\psi^*)$ are computable. If $C<T$, $U(\psi^*)$ is not computable; however, in this case, we shall show that $C(\psi^*)<U(\psi^*)$ corresponding to $\widetilde{U}(\psi^*)=C(\psi^*)$ and $\Delta(\psi^*)=0$, which are computable. Toward this end, by definition of $C(\psi^*)$, we show that when $C<T$, it is always the case that $C(\psi^*)\leqslant U(\psi^*)$. If $\psi^*\geqslant 0$, $C(\psi^*)=C\leqslant T\leqslant\int_0^T\exp(\psi^*A_u)\mathrm{d}u=U(\psi^*)$. If $\psi^*<0$, $C(\psi^*)=C\exp(\psi^*)\leqslant T\exp(\psi^*)=\int_0^T\exp(\psi^*)\mathrm{d}u\leqslant\int_0^T\exp(\psi^*A_u)\mathrm{d}u=U(\psi^*)$. This completes the proof.

S8. Proof of $\lambda_V(t\mid \overline{H}_t)=\lambda_V(t\mid \overline{H}_t, C\geqslant t)$

First, by Assumption 1, we obtain

$$
\begin{aligned}
P(C\geq t\mid t\leqslant V<t+h,\Gamma=1,\overline{H}_t)&=\exp\left\{\int_0^t-\lambda_C(u\mid t\leqslant V<t+h,\Gamma=1,\overline{H}_t)\mathrm{d}u\right\}\\
&=\exp\left\{\int_0^t-\lambda_C(u\mid \overline{H}_u)\mathrm{d}u\right\},
\end{aligned}
$$

and similarly, we obtain

$$
\begin{aligned}
P(C\geq t\mid V\geq t,\Gamma=1,\overline{H}_t)&=\exp\left\{\int_0^t-\lambda_C(u\mid V\geq t,\Gamma=1,\overline{H}_t)\mathrm{d}u\right\}\\
&=\exp\left\{\int_0^t-\lambda_C(u\mid \overline{H}_u)\mathrm{d}u\right\}.
\end{aligned}
$$

Consequently, $P(C\geq t\mid t\leqslant V<t+h,\Gamma=1,\overline{H}_t)=P(C\geq t\mid V\geq t,\Gamma=1,\overline{H}_t)$.

Now, by the Bayes rule, we express

$$
\lambda_V(t\mid \overline{H}_t, C\geqslant t)=\lim_{h\to 0}h^{-1}P(t\leqslant V<t+h,\Gamma=1\mid V\geqslant t,\overline{H}_t, C\geq t)
$$

$$= \lim_{h\to 0} h^{-1} \frac{P(t \leqslant V < t+h, \Gamma = 1 \mid V \geqslant t, \overline{H}_t) P(C \geq t \mid t \leqslant V < t+h, \Gamma = 1, \overline{H}_t)}{P(C \geq t \mid V \geq t, \overline{H}_t)}$$
$$= \lim_{h\to 0} h^{-1} P(t \leqslant V < t+h, \Gamma = 1 \mid V \geqslant t, \overline{H}_t) = \lambda_V(t \mid \overline{H}_t).$$

S9. IDENTIFICATION OF $\psi \in \mathcal{R}^p$ UNDER ASSUMPTIONS 1–3

Under Assumptions 1–3, for any $c(\overline{H}_t) \in \mathcal{R}^p$ and $t > 0$,

$$E\left\{ \frac{\Delta}{K_C\left(T \mid \overline{H}_T\right)} c(\overline{H}_t) U(\psi^*) \mathrm{d}M_V(t) \right\} = E\left\{ c(\overline{H}_t) U(\psi^*) \mathrm{d}M_V(t) \right\} = 0. \tag{S14}$$

Because under Assumption 1, ψ^* is uniquely identified from (S2). Therefore, under Assumptions 1–3, ψ^* is uniquely identified from (S14).

S10. PROOF OF THEOREM 3

To show (6) is an unbiased estimating equation, it suffices to show that

$$E\left\{ \frac{\Delta}{K_C\left(T \mid \overline{H}_T\right)} G(\psi^*; F) \right\} = 0.$$

Toward that end, by the iterative expectation, we have

$$\begin{aligned} E\left\{ \frac{\Delta}{K_C\left(T \mid \overline{H}_T\right)} G(\psi^*; F) \right\} &= E\left[E\left\{ \frac{\Delta}{K_C\left(T \mid \overline{H}_T\right)} G(\psi^*; F) \mid F \right\} \right] \\ &= E\left\{ \frac{E(\Delta \mid F)}{K_C\left(T \mid \overline{H}_T\right)} G(\psi^*; F) \right\} \\ &= E\left\{ 1 \times G(\psi^*; F) \right\} = 0, \end{aligned}$$

where the third equality follows by the dependent censoring mechanism specified in Assumption 2.

S11. THE ASYMPTOTIC PROPERTIES OF THE PROPOSED ESTIMATOR

To establish the asymptotic properties of the proposed estimator, we first introduce additional notation.

Recall the nuisance models (i) $E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi\}$ indexed by ξ; (ii) the proportional hazards model for the treatment process, indexed by M_V; and (iii) the proportional hazards model for the censoring process, indexed by K_C. $\widehat{\xi}$, $\widehat{M}_V$, and $\widehat{K}_C$ are the estimates of ξ, M_V, and K_C under the specified parametric and semiparametric models. The probability limits of $\widehat{\xi}$, $\widehat{M}_V$, and $\widehat{K}_C$ are ξ^*, M_V^*, and K_C^*. If the failure time model is correctly specified, $E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi^*\} = E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\}$; if the model for the treatment process is correctly specified, $M_V^* = M_V$; and if the model for the censoring process is correctly specified, $K_C^* = K_C$.

To reflect that the estimating function depends on the nuisance parameters, we define

$$\Phi(\psi, \xi, M_V, K_C; F) = \frac{\Delta G(\psi, \xi, M_V; F)}{K_C\left(T \mid \overline{H}_T\right)},$$
$$G(\psi, \xi, M_V; F) = \int c(\overline{H}_u)\left[U(\psi) - E\left\{U(\psi) \mid \overline{H}_u, V \geqslant u; \xi\right\}\right] \mathrm{d}M_V(t).$$

Let P denote the true data generating distribution, and for any $f(F)$, let $P\{f(F)\} = \int f(x)\mathrm{d}P(x)$. We define

$$J_1(\xi) = P\left\{\Phi(\psi^*, \xi, M_V^*, K_C^*; F)\right\},$$
$$J_2(M_V) = P\left\{\Phi(\psi^*, \xi^*, M_V, K_C^*; F)\right\},$$
$$J_3(K_C) = P\left\{\Phi(\psi^*, \xi^*, M_V^*, K_C; F)\right\},$$

and

$$J(\xi, M_V, K_C) = P\left\{\Phi(\psi^*, \xi, M_V, K_C; F)\right\}.$$

We now assume the regularity conditions, which are standard in the empirical process literature (van der Vaart & Wellner, 1996). See also Yang & Lok (2016) for the application of the empirical process to derive a goodness-of-fit test for the structural nested mean models.

*Assumption S*1. With probability going to one, $\Phi(\psi, \xi, M_V, K_C; F)$ and $\partial\Phi(\psi, \xi, M_V, K_C; F)/\partial\psi$ are P-Donsker classes.

*Assumption S*2. For (ξ^*, M_V^*, K_C^*) with either ξ^* being the true parameter such that $E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi^*\} = E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\}$ or $M_V^* = M_V$, and $K_C^* = K_C$,

$$P\left\{||\Phi(\psi^*, \widehat{\xi}, \widehat{M}_V, \widehat{K}_C; F) - \Phi(\psi^*, \xi^*, M_V^*, K_C^*; F)||\right\} \to 0$$

and

$$P\left\{||\frac{\partial}{\partial\psi}\Phi(\widehat{\psi}, \widehat{\xi}, \widehat{M}_V, \widehat{K}_C; F) - \frac{\partial}{\partial\psi}\Phi(\psi^*, \xi^*, M_V^*, K_C^*; F)||\right\} \to 0$$

in probability.

*Assumption S*3. $A(\psi^*, \xi^*, M_V^*, K_C^*) = P\left\{\partial\Phi(\psi^*, \xi^*, M_V^*, K_C^*; F)/\partial\psi\right\}$ is invertible.

*Assumption S*4. Assume that

$$J(\widehat{\xi}, \widehat{M}_V, \widehat{K}_C) - J(\xi^*, M_V^*, K_C^*) = J_1(\widehat{\xi}) - J_1(\xi^*) + J_2(\widehat{M}_V) - J_2(M_V^*) + J_3(\widehat{K}_C) - J_3(K_C^*) + o_p(n^{-1/2}),$$

and that $J_1(\widehat{\xi})$, $J_2(\widehat{M}_V)$, and $J_3(\widehat{K}_C)$ are regular asymptotically linear with influence function $\Phi_1(\psi^*, \xi^*, M_V^*, K_C^*; F)$, $\Phi_2(\psi^*, \xi^*, M_V^*, K_C^*; F)$, and $\Phi_3(\psi^*, \xi^*, M_V^*, K_C^*; F)$, respectively.

Assumption S1 is an empirical process condition. This assumption is technical and depends on the submodel chosen models for the unknown parameters. Assuming the positivity condition for the censoring process, this assumption can typically be considered as a regularity condition.

Assumption S2 basically states that $\widehat{\xi}$, $\widehat{M}_V$, and $\widehat{K}_C$ are consistent for ξ^*, M_V^*, and K_C and requires that

$$E\left\{\int c(\overline{H}_u)\left[E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \widehat{\xi}\right\}\right.\right.$$
$$\left.\left. - E\left\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi^*\right\}\right]\left\{\widehat{\lambda}_V(u) - \lambda_V^*(u)\right\} \mathrm{d}u = o(n^{-1/2}),$$

and

$$E\left\{\int c(\overline{H}_u)\left[E\left\{\frac{\partial U(\psi^*)}{\partial \psi} \mid \overline{H}_u, V \geqslant u; \widehat{\xi}\right\}\right.\right.$$
$$\left.\left. - E\left\{\frac{\partial U(\psi^*)}{\partial \psi} \mid \overline{H}_u, V \geqslant u; \xi^*\right\}\right]\left\{\widehat{\lambda}_V(u) - \lambda_V^*(u)\right\} \mathrm{d}u\right\} = o(n^{-1/2}).$$

Because smooth functionals of parametric or semiparametric maximum likelihood estimators for a given model are efficient under regularity conditions, Assumption S4 holds under regularity conditions if $\widehat{\xi}$ and $\widehat{M}_V$ are the parametric and semiparametric maximum likelihood estimators of ξ^* and M_V^* under the specified models.

We present the asymptotic properties of the proposed estimator $\widehat{\psi}$ solving (7), denoted by $P_n\left\{\Phi(\psi, \widehat{\xi}, \widehat{M}_V, \widehat{K}_C; F)\right\} = 0$.

THEOREM S3. *Under Assumptions 3 and S1–S4, $n^{1/2}\left(\widehat{\psi} - \psi^*\right)$ is consistent and asymptotically linear with the influence function $\widetilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F) = \{A(\psi^*, \xi^*, M_V^*, K_C^*)\}^{-1} \widetilde{B}(\psi^*, \xi^*, M_V^*, K_C^*; F)$, and*

$$\widetilde{B}(\psi^*, \xi^*, M_V^*, K_C^*; F) = \Phi(\psi^*, \xi^*, K_V^*, K_C^*; F) + \Phi_1(\psi^*, \xi^*, K_V^*, K_C^*; F)$$
$$+\Phi_2(\psi^*, \xi^*, K_V^*, K_C^*; F) + \Phi_3(\psi^*, \xi^*, K_V^*, K_C^*; F). \quad \text{(S15)}$$

Moreover, if the nuisance models including the models for the censoring process and the treatment process and the outcome model are correctly specified, (S15) becomes

$$\begin{aligned}
&\widetilde{B}(\psi^*, \xi^*, K_V, K_C; F) \\
&= \Phi(\psi^*, \xi^*, K_V, K_C; F) - \prod\left\{\Phi(\psi^*, \xi^*, K_V, K_C; F) \mid \widetilde{\Lambda}\right\} \\
&= \Phi(\psi^*, \xi^*, K_V, K_C; F) - E\left\{\Phi(\psi^*, \xi^*, K_V, K_C; F) S_{\gamma_V}^{\mathrm{T}}\right\} E\left(S_{\gamma_V} S_{\gamma_V}^{\mathrm{T}}\right)^{-1} S_{\gamma_V} \\
&\quad - E\left\{\Phi(\psi^*, \xi^*, K_V, K_C; F) S_{\gamma_C}^{\mathrm{T}}\right\} E\left(S_{\gamma_C} S_{\gamma_C}^{\mathrm{T}}\right)^{-1} S_{\gamma_C} \\
&\quad + \int \frac{E\left[G(\psi^*, \xi^*, K_V; F) \exp\left\{\gamma_C^{\mathrm{T}} g_C(u, \overline{H}_u)\right\} \Delta / K_C(T \mid \overline{H}_T)\right]}{E\left[\exp\left\{\gamma_C^{\mathrm{T}} g_C(u, \overline{H}_u)\right\} Y_C(u)\right]} \mathrm{d}M_C(u) \\
&\quad + \int \frac{E\left[G(\psi^*, \xi^*, K_V; F) \exp\left\{\gamma_V^{\mathrm{T}} g_V(u, \overline{H}_u)\right\} \Delta / K_C(T \mid \overline{H}_T)\right]}{E\left[\exp\left\{\gamma_V^{\mathrm{T}} g_V(u, \overline{H}_u)\right\} Y_V(u)\right]} \mathrm{d}M_V(u). \quad \text{(S16)}
\end{aligned}$$

Proof. We assume that the model for the censoring process is correctly specified, either the outcome model or the model for the treatment process is correctly specified.

Taylor expansion of $P_n\left\{\Phi(\widehat{\psi},\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)\right\}=0$ around ψ^* leads to

$$0=P_n\left\{\Phi(\widehat{\psi},\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)\right\}=P_n\left\{\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)\right\} + P_n\left\{\frac{\partial\Phi(\widetilde{\psi},\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)}{\partial\psi^{\mathrm{T}}}\right\}(\widehat{\psi}-\psi^*),$$

where $\widetilde{\psi}$ is on the line segment between $\widehat{\psi}$ and ψ^*.

Under Assumptions S1 and S2,

$$(P_n-P)\left\{\frac{\partial\Phi(\widetilde{\psi},\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)}{\partial\psi^{\mathrm{T}}}\right\}=(P_n-P)\left\{\frac{\partial\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F)}{\partial\psi^{\mathrm{T}}}\right\}=o_p(n^{-1/2}),$$

and therefore,

$$P_n\left\{\frac{\partial\Phi(\widetilde{\psi},\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)}{\partial\psi^{\mathrm{T}}}\right\}=P\left\{\frac{\partial\Phi(\widetilde{\psi},\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)}{\partial\psi^{\mathrm{T}}}\right\}+o_p(n^{-1/2}) =A(\psi^*,\xi^*,M_V^*,K_C^*)+o_p(n^{-1/2}).$$

We then have

$$n^{1/2}(\widehat{\psi}-\psi^*)=\{A(\psi^*,\xi^*,M_V^*,K_C^*)\}^{-1}\,n^{1/2}P_n\left\{\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)\right\}+o_p(1). \quad \text{(S17)}$$

To evaluate (S17) further,

$$P_n\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)=(P_n-P)\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F) + P\left\{\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)-\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F)\right\}+P\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F). \quad \text{(S18)}$$

Based on the double robustness, the third term becomes

$$P\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F)=0. \quad \text{(S19)}$$

By Assumptions S1 and S2, the first term becomes

$$(P_n-P)\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)=(P_n-P)\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F)+o_p(n^{-1/2}) =P_n\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F)+o_p(n^{-1/2}). \quad \text{(S20)}$$

By Assumption S4, the second term becomes

$$\begin{aligned}&P\left\{\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)-\Phi(\psi^*,\xi^*,M_V^*,K_C^*;F)\right\}\\ &=J(\widehat{\xi},\widehat{M}_V,\widehat{K}_C)-J(\xi^*,M_V^*,K_C^*)+o_p(n^{-1/2})\\ &=J_1(\widehat{\xi})-J_1(\xi^1)+J_2(\widehat{M}_V)-J_2(M_V^1)+J_3(\widehat{K}_C)-J_3(M_C^1)+o_p(n^{-1/2})\\ &=P_n\left\{\Phi_1(\psi^*,\xi^*,M_V^*,K_C^*;F)+\Phi_2(\psi^*,\xi^*,M_V^*,K_C^*;F)+\Phi_3(\psi^*,\xi^*,M_V^*,K_C^*;F)\right\}\end{aligned} \quad \text{(S21)}$$

Combining (S20)–(S19) with (S18),

$$P_n\Phi(\psi^*,\widehat{\xi},\widehat{M}_V,\widehat{K}_C;F)=P_n\{\widetilde{\Phi}(\psi^*,\xi^*,M_V^*,K_C^*;F)\},$$

where

$$\widetilde{B}(\psi^*, \xi^*, M_V^*, K_C^*; F) = \Phi(\psi^*, \xi^*, M_V^*, K_C^*; F) + \Phi_1(\psi^*, \xi^*, M_V^*, K_C^*; F) + \Phi_2(\psi^*, \xi^*, M_V^*, K_C^*; F) + \Phi_3(\psi^*, \xi^*, M_V^*, K_C^*; F).$$

Therefore, $\widehat{\psi} - \psi^*$ has the influence function

$$\widetilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F) = \{A(\psi^*, \xi^*, M_V^*, K_C^*)\}^{-1} \widetilde{B}(\psi^*, \xi^*, M_V^*, K_C^*; F).$$

As a result,

$$n^{1/2}(\widehat{\psi} - \psi^*) = n^{1/2} P_n \widetilde{\Phi}(\psi^*, \xi^*, K_V^*, K_C^*; F) + o_p(1). \tag{S22}$$

Based on (S22),

$$n^{1/2}(\widehat{\psi} - \psi^*) \to \mathcal{N}(0, \Omega),$$

as $n \to \infty$, where $\Omega = E\left\{\widetilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F)\widetilde{\Phi}(\psi^*, \xi^*, M_V^*, K_C^*; F)^{\mathrm{T}}\right\}$.

For the special case where both nuisance models are correctly specified, we characterize $\widetilde{B}(\psi^*, \xi^*, K_V^*, K_C^*; F)$. In this case, $E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u; \xi^*\} = E\{U(\psi^*) \mid \overline{H}_u, V \geqslant u\}$, $M_V^* = M_V$, and $K_C^* = K_C$. Define the score functions: $S_\xi = S_\xi\{U(\psi^*), \overline{H}_u, V \geqslant u\}$,

$$S_{\gamma_V} = \int \left\{ g_V(u, \overline{H}_u) - \frac{E\left[g_V(u, \overline{H}_u) \exp\left\{\gamma_V^{\mathrm{T}} g_V(u, \overline{H}_u)\right\} Y_V(u)\right]}{E\left[\exp\left\{\gamma_V^{\mathrm{T}} g_V(u, \overline{H}_u)\right\} Y_V(u)\right]} \right\} \mathrm{d}M_V(u),$$

and

$$S_{\gamma_C} = \int \left\{ g_C(u, \overline{H}_u) - \frac{E\left[g_C(u, \overline{H}_u) \exp\left\{\gamma_C^{\mathrm{T}} g_C(u, \overline{H}_u)\right\} Y_C(u)\right]}{E\left[\exp\left\{\gamma_C^{\mathrm{T}} g_C(u, \overline{H}_u)\right\} Y_C(u)\right]} \right\} \mathrm{d}M_C(u).$$

The tangent space for ξ is $\widetilde{\Lambda}_1 = \{S_\xi \in \mathcal{R}^p : E(S_\xi \mid \overline{H}_u, V \geqslant u) = 0\}$. Following Tsiatis (2006), the nuisance tangent space for the proportional hazards model of the treatment process is

$$\widetilde{\Lambda}_2 = \left\{ S_{\gamma_V} + \int h(u) \mathrm{d}M_V(u) : \ h(u) \in \mathcal{R}^p \right\},$$

and the nuisance tangent space for the proportional hazards model of the censoring process is

$$\widetilde{\Lambda}_3 = \left\{ S_{\gamma_C} + \int h(u) \mathrm{d}M_C(u) : \ h(u) \in \mathcal{R}^p \right\}.$$

Assuming that the treatment process and the censoring process can not jump at the same time point, $\widetilde{\Lambda}_1$, $\widetilde{\Lambda}_2$, and $\widetilde{\Lambda}_3$ are mutually orthogonal to each other. Therefore, the nuisance tangent space for ξ and the proportional hazards models is $\widetilde{\Lambda} = \widetilde{\Lambda}_1 \oplus \widetilde{\Lambda}_2 \oplus \widetilde{\Lambda}_3$. The influence function

for $\widehat{\psi}$ is

$$
\begin{aligned}
&\widetilde{B}(\psi^*, \xi^*, M_V, K_C; F) \\
&= \Phi(\psi^*, \xi^*, M_V, K_C; F) - \prod \left\{ \Phi(\psi^*, \xi^*, M_V, K_C; F) \mid \widetilde{\Lambda} \right\} \\
&= \Phi(\psi^*, \xi^*, M_V, K_C; F) - E\left\{ \Phi(\psi^*, \xi^*, M_V, K_C; F) S_{\gamma_V}^{\mathrm{T}} \right\} E\left(S_{\gamma_V} S_{\gamma_V}^{\mathrm{T}} \right)^{-1} S_{\gamma_V} \\
&\quad - E\left\{ \Phi(\psi^*, \xi^*, M_V, K_C; F) S_{\gamma_C}^{\mathrm{T}} \right\} E\left(S_{\gamma_C} S_{\gamma_C}^{\mathrm{T}} \right)^{-1} S_{\gamma_C} \\
&\quad + \int \frac{E\left[G(\psi^*, \xi^*, M_V; F) \exp\left\{ \gamma_C^{\mathrm{T}} g_C(u, \overline{H}_u) \right\} \Delta / K_C(T \mid \overline{H}_T) \right]}{E\left[\exp\left\{ \gamma_C^{\mathrm{T}} g_C(u, \overline{H}_u) \right\} Y_C(u) \right]} \mathrm{d}M_C(u) \\
&\quad + \int \frac{E\left[G(\psi^*, \xi^*, M_V; F) \exp\left\{ \gamma_V^{\mathrm{T}} g_V(u, \overline{H}_u) \right\} \Delta / K_C(T \mid \overline{H}_T) \right]}{E\left[\exp\left\{ \gamma_V^{\mathrm{T}} g_V(u, \overline{H}_u) \right\} Y_V(u) \right]} \mathrm{d}M_V(u). \qquad \square
\end{aligned}
$$

S12. The Cox marginal structural model approach: $\widehat{\psi}_{\mathrm{msm}}$

The Cox marginal structural model approach assumes that the potential failure time under $\overline{a}_T$ follows a Cox proportional hazards model with the hazard rate at t as $\lambda_0(t) \exp(\psi^* a_t)$.

If all potential failure times were observed for all subjects, one can fit a Cox proportional hazards model with the time-varying covariate a_t to obtain a consistent estimator of ψ^*. However, not all potential outcomes are observed for a particular subject. To obtain a consistent estimator based on the actual observed data, the key step is to construct time-dependent inverse probability of treatment weights for all subjects and weight their contributions so that they mimic the contributions had all potential outcomes been observed.

From the hazard of treatment discontinuation $\lambda_V\left(t \mid \overline{H}_t\right)$ defined in (2), denote

$$
K_V\left(t \mid \overline{H}_t\right) = \exp\left\{ -\int_0^t \lambda_V(u \mid \overline{H}_u) \mathrm{d}u \right\} \tag{S23}
$$

and

$$
f_V\left(t \mid \overline{H}_t\right) = \lambda_V\left(t \mid \overline{H}_t\right) K_V\left(t \mid \overline{H}_t\right). \tag{S24}
$$

For ease of notation, denote $K_V(t) = K_V\left(t \mid \overline{H}_t\right)$ and $f_V(t) = f_V\left(t \mid \overline{H}_t\right)$ for shorthand. These can be viewed as the probability of not having discontinued before time t and the probability of discontinuing at time $[t, t + \mathrm{d}t)$, respectively.

Consider subjects who were are at risk at time t. We consider two subsets of individuals: group (a) with $V \leqslant t$ and $\Gamma = 1$ and group (b) with $V > t$. Specifically, we construct the time-dependent inverse probability of treatment weight as

$$
\omega(t) = \begin{cases} \theta(V)/f_V(V), & \text{if } V \leqslant t \text{ and } \Gamma = 1, \\ \overline{\theta}(t)/K_V(t) & \text{if } V > t, \end{cases} \tag{S25}
$$

where $\theta(t)$ and $\overline{\theta}(t) = \int_t^\infty \theta(u)\mathrm{d}u$ serve as the stabilized weights (Hernán et al., 2000). Following (Yang et al., 2018), one can consider $\theta(t) = \lambda_{V,0}(t) \exp\left\{ -\int_0^t \lambda_{V,0}(u)\mathrm{d}u \right\}$. In the presence of censoring, let $\omega(t)$ be a product of (S25) and the inverse of censoring probability $\Delta / K_C\left(T \mid \overline{H}_T\right)$. One can estimate the weights by replacing the unknown quantities with their estimates following Steps 1 and 2 in § 4.

Finally, we obtain $\widehat{\psi}_{\text{msm}}$ by fitting a Cox proportional hazards model with the time-varying covariate A_t with the time-dependent weight $\omega(t)$ using the standard software; e.g., the function "coxph" in R with the weighting argument.

S13. THE DISCRETE-TIME G-ESTIMATOR: $\widehat{\psi}_{\text{disc}}$

The existing framework for fitting the structural failure time model is using a discrete time points setting which requires manually discretizing the data. We disretize the timeline into equally-spaced time points from 0 to the maximum follow up τ, denoted as $0 = t_0 < t_1 < \cdots < t_K = \tau$. For $m \geqslant 1$, at the mth time point t_m, let A_{t_m} be the indicator of whether the treatment is received at t_m, let L_{t_m} be the the average of L_t from $t_{m-1} \leqslant t \leqslant t_m$, let H_{t_m} be the vector of A_{t_m-1} and L_{t_m}, and finally let $\overline{H}_{t_m}$ be $\{H_0, \ldots, H_{t_m}\}$. With observations at discrete time points, $\mathrm{d}N_T(t_m)$ becomes the binary treatment indicator A_{t_m}, $\lambda_T(u \mid \overline{H}_u)Y_T(u)\mathrm{d}u$ becomes the propensity score $E(A_{t_m} \mid \overline{H}_{t_m}, \overline{A}_{t_m-1} = \overline{0})$, and the integral in (3) becomes the summation from $m = 1$ to K. As a result, in the absence of censoring, (4) simplifies to the existing estimating equation for structural nested failure time models (Hernán et al., 2005). Following (Hernán et al., 2005), one can estimate the propensity score by the pooled logistic regression model with baseline and time-dependent covariates. In the presence of censoring, one can estimate the censoring probability by the pooled logistic regression model with baseline and time-dependent. The g-estimator $\widehat{\psi}_{\text{disc}}$ of ψ^* solves estimating equation (7) with observations at discrete time points.

S14. DETAILS AND ADDITIONAL RESULTS IN THE SIMULATION

In this section, we present details for the Jackknife method for variance estimation and additional simulation results to assess the impact of misspecification of the censoring model and the treatment effect model.

The Jackknife method entails dividing the subjects into exclusive and exhaustive subgroups, creating replicate datasets by deleting one group at a time, and applying the same estimation procedure to obtain the replicates of $\widehat{\psi}$. The variance estimator is $\widehat{V}(\widehat{\psi}) = G^{-1}(G-1)\sum_{k=1}^{G}\left(\widehat{\psi}^{(k)} - \widehat{\psi}\right)^2$, where G is the number of subgroups, and $\widehat{\psi}^{(k)}$ is the kth replicate of $\widehat{\psi}$.

We now focus on the scenario 1 of the simulation study in § 5. First in setting 1, to illustrate the impact of misspecification of the censoring model, for all estimators, we consider an incorrect independent censoring mechanism for fitting the censoring model in the sense that the censoring indicator is independent of all other variables. Second in setting 2, to illustrate the impact of misspecification of the treatment effect model, we now generate the failure time, T, according to a structural failure time model $U \sim \int_0^T \exp(\psi^* A_u + 0.5X_0)\mathrm{d}u$. All estimators are the same as in § 5.

Table S1 summarizes the simulation results with $n = 1,000$. In setting 1 when the censoring model is misspecified, the proposed estimators have larger biases compared to the results when the censoring model is correctly specified as in Table 1. In setting 2 when the treatment effect model is misspecified, the proposed estimators also have increased biases compared to the results when the treatment effect model is correctly specified as in Table 1. The coverage rates are off the nominal coverage in most of cases.

Table S1. *Simulation results: bias, standard deviation, root mean squared error, and coverage rate of* 95% *confidence intervals for* $\exp(\psi^*)$ *over* $1,000$ *simulated datasets: Setting 1 where the censoring model is misspecified, and Setting 2 where the treatment effect model is misspecified*

			$\psi^* = -0.5$			$\psi^* = 0$			$\psi^* = 0.5$		
			Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.
Setting 1	$\widehat{\psi}_{\text{naive}}$	c	0.06	0.048	76.8	0.02	0.069	95.6	-0.06	0.112	92.4
		c^{opt}	0.05	0.043	78.4	0.02	0.063	95.0	-0.05	0.107	91.8
	$\widehat{\psi}_{\text{ipcw}}$	c	0.14	0.086	68.2	0.04	0.103	97.4	-0.13	0.153	82.2
		c^{opt}	0.13	0.072	61.6	0.04	0.089	96.8	-0.12	0.136	81.4
	$\widehat{\psi}_{\text{dr}}$	c	0.04	0.050	88.4	0.01	0.070	95.0	-0.04	0.116	94.0
		c^{opt}	0.04	0.048	89.0	0.01	0.069	95.2	-0.03	0.115	93.2
	$\widehat{\psi}_{\text{msm}}$	–	0.04	0.045	92.6	0.02	0.073	96.8	-0.04	0.135	94.6
	$\widehat{\psi}_{\text{disc}}$	–	-0.40	0.035	0.0	-0.65	0.046	0.0	-1.09	0.070	0.0
Setting 2	$\widehat{\psi}_{\text{naive}}$	c	0.05	0.049	87.0	-0.01	0.069	94.8	-0.12	0.108	80.2
		c^{opt}	0.03	0.044	90.8	-0.03	0.064	92.2	-0.14	0.100	73.4
	$\widehat{\psi}_{\text{ipcw}}$	c	-0.02	0.088	95.6	-0.05	0.120	94.6	-0.10	0.174	90.8
		c^{opt}	-0.04	0.070	93.4	-0.07	0.096	93.2	-0.12	0.136	86.2
	$\widehat{\psi}_{\text{dr}}$	c	-0.02	0.053	94.2	-0.03	0.079	91.0	-0.08	0.122	90.6
		c^{opt}	-0.02	0.050	91.6	-0.05	0.073	89.4	-0.11	0.115	84.8
	$\widehat{\psi}_{\text{msm}}$	–	-0.01	0.049	96.6	-0.03	0.082	92.2	-0.09	0.134	91.6
	$\widehat{\psi}_{\text{disc}}$	–	-0.38	0.041	0.0	-0.63	0.053	0.0	-1.06	0.085	0.2

S15. NUISANCE MODELS IN THE APPLICATION

In this section, we provide details for fitting the nuisance models in the application. To build a model for $\lambda_V(t \mid \overline{H}_t)$, we consider the baseline covariates X, including age, gender, race, site, country, and other 25 baseline health outcome measures. For each categorical variable, we create dummy variables. This leads to 99 baseline variables. We first fit a Cox proportional hazards model for $\lambda_V(t \mid \overline{H}_t)$ to the data including the baseline variables with a l_1 penalty. In fitting the model, we select the tuning parameter using 10-fold cross-validation. The final proportional hazards model includes the selected baseline terms and all time-dependent covariates L_t, including indicators of bleeding, haemorrhagic stroke, and left atrial appendage procedures associated with permanent discontinuation and outcomes. To build a model for $\lambda_C(t \mid \overline{H}_t)$, we consider the same procedure for $\lambda_V(t \mid \overline{H}_t)$. This is because the decision to re-start treatment was left to the patient and physician, and the resulting censoring may depend on the patient's characteristics and evolving disease status. To estimate $E\left\{U(\psi) \mid \overline{H}_0\right\}$, we regress $\widehat{K}_C\left(T \mid \overline{H}_T\right)^{-1} \Delta U(\psi)$ on X with a l_1 penalty.

REFERENCES

ANDERSEN, P. K., BORGAN, O., GILL, R. D. & KEIDING, N. (1993). *Statistical Models based on Counting Processes*. Springer-Verlag, New York.

HERNÁN, M. Á., BRUMBACK, B. & ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11**, 561–570.

HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. & ROBINS, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* **14**, 477–491.

TSIATIS, A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Emprical Processes: With Applications to Statistics*. New York: Springer.

YANG, S. & LOK, J. J. (2016). A goodness-of-fit test for structural nested mean models. *Biometrika* **103**, 734–741.

YANG, S., TSIATIS, A. A. & BLAZING, M. (2018). Modeling survival distribution as a function of time to treatment discontinuation: A dynamic treatment regime approach. *Biometrics* **74**, 900–909.